Scientometrics
https://doi.org/10.1007/s11192-018-2803-x

LitStoryTeller+: an interactive system for multi-level scientific paper visual storytelling with a supportive text mining toolbox

Qing Ping[1] · **Chaomei Chen**[1]

Received: 31 January 2018

Abstract The continuing growth of scientific publications has posed a double-challenge to researchers, to not only grasp the overall research trends in a scientific domain, but also get down to research details embedded in a collection of core papers. Existing work on science mapping provides multiple tools to visualize research trends in domain on macro-level, and work from the digital humanities have proposed text visualization of documents, topics, sentences, and words on micro-level. However, existing micro-level text visualizations are not tailored for scientific paper corpus, and cannot support meso-level scientific reading, which aligns a set of core papers based on their research progress, before drilling down to individual papers. To bridge this gap, the present paper proposes LitStoryTeller+, an interactive system under a unified framework that can support both meso-level and micro-level scientific paper visual storytelling. More specifically, we use entities (concepts and terminologies) as basic visual elements, and visualize entity storylines across papers and within a paper borrowing metaphors from screen play. To identify entities and entity communities, named entity recognition and community detection are performed. We also employ a variety of text mining methods such as extractive text summarization and comparative sentence classification to provide rich textual information supplementary to our visualizations. We also propose a top-down story-reading strategy that best takes advantage of our system. Two comprehensive hypothetical walkthroughs to explore documents from the computer science domain and history domain with our system demonstrate the effectiveness of our story-reading strategy and the usefulness of LitStoryTeller+.

Keywords Visual storytelling · Narrative storylines · Close-reading · Scientific paper visualization · Extractive summarization · Comparative sentence classification

The present study is an extended version of an article (Ping and Chen 2017) presented at the 16th International Conference on Scientometrics and Informetrics, Wuhan (China), 16–20 October 2017.

✉ Qing Ping
qp27@drexel.edu

[1] College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

Published online: 11 June 2018

Introduction

It is estimated that the growth rate of new scientific publications is close to 8–9% each year, leading to a doubling of global scientific output roughly every nine years (Bornmann and Mutz 2015). This has become a double-challenge for researchers who want to keep up with the research trends, and develop novel research ideas. On the one hand, researchers need to have a macro-level picture of the discipline, in terms of what are the new research trends, what are the different sub-fields, what are the interactions between these sub-fields, and so on. On the other hand, researchers must keep in mind the research details on meso-level and micro-level by reading scientific publications. Meso-level research details refer to the hidden pattern of how different approaches have been applied on a very specific research problem progressively, each with its own novel contribution. Micro-level research details refer to knowledge of how and why a new method works better than old methods. This core principle is derived by author's logical thinking process and manifested in his or her scientific writing.

To tackle this double-challenge, on the macro-level, the science-mapping community has proposed multiple visualization applications that can help users to get an overall picture of the entire research domain, such as CiteSpace (Chen 2006), VOSViewer (Van Eck and Waltman 2010), Action Science Explorer (Dunne et al. 2012) and so on. On the micro-level, in the domain of digital humanities, several applications have been developed to facilitate document digestion on topic-level, such as VarifocalReader (Koch et al. 2014), Serendip (Alexander et al. 2014), on sentence-level, such as PICTOR (Schneider et al. 2010), and on word-level, such as POSvis (Clement et al. 2009) and Wordle (Viegas et al. 2009).

However, existing research is insufficient in resolving the double-challenge. While science-mapping can draw a global picture of a research domain on macro-level, visualization tools to support scientific paper reading on meso-level and micro-level are still rare if not none. Existing work in digital humanities usually focus on visualizing specific types of corpus, such as poem, play, news, Bible, and so on, but very few are tailored for scientific papers. Nevertheless, on meso and micro-level, it is still a challenge for a reader to quickly disentangle the intricate relationships between different methods in different papers, and get all the important details in each individual scientific paper.

To bridge this gap, we propose LitStoryTeller+, an interactive system that can support multi-level scientific visual storytelling, with a supportive text-mining toolbox. In our previous work (Ping and Chen 2017), we have proposed LitStoryTeller, an interactive system that supports visual storytelling of a single scientific paper at micro-level. In the present paper, we incorporate LitStoryTeller into our unified system, to support both meso-level and micro-level scientific paper storytelling, namely to visualize both storylines of multi-documents on the time axis, and storylines of single-document on the narrative axis. More specifically, we use entities, or concepts and terminologies in research papers, as the basic visual elements in our system. We borrow metaphor from screen play, treating entities as characters, paper/paragraph/sentence as scenes that characters co-stage, and visualize the storylines of entities across papers/paragraphs/sentences. We utilize named entity recognition and community detection techniques to identify entities and their "co-stage" associations from full-text of research papers. Besides these visual storylines, we also employ a variety of text mining techniques, including text summarization and comparative sentence classification to provide supportive contextual information supplementary to our visual storylines.

To our best knowledge, this paper is among the first work that is designed to support scientific paper reading at both meso-level and micro-level using a storyline visual metaphor and leveraging a variety of text mining techniques. The main contributions of the present work are as follows:

1. We propose a unified visualization framework for scientific paper storytelling, using metaphors from screen play to draw entity storylines of multiple papers over time axis, and entity storylines of single paper over narrative axis;
2. We build a supportive text-mining toolbox that could perform named entity recognition, community detection, text summarization and comparative sentence classification on the fly for arbitrary scientific papers;
3. We propose a top-down story-reading strategy that starts from reading an overall entity co-occurrence network, to reading multi-document storylines, to reading single-document storylines. Two hypothetical walkthroughs on different topics from different domains demonstrated the effectiveness of this top-down story-reading strategy.

The rest of the paper is organized as follows. In the Section "Related work", we review existing work on two areas, namely multiple/single document visualization, and natural language processing algorithms relevant to our system. In the Section "System pipeline", we give an overview of our system by describing the system pipeline, and the functions of each component in the pipeline. In the Section "System components", we describe each component in detail, starting from pre-processing, to "back-end" natural-language-processing, to "front-end" visualization. In the Section "Storyline-reading strategy", we propose a view of our system at a more abstract level, and propose a serial reading strategy that is hierarchical in nature. In the section "Hypothetical Walkthrough-I" and "Hypothetical Walkthrough-II", we demonstrate the use of our system in two concrete and complete cases, utilizing documents of different research topics from different domains. In the Section "Conclusion", we draw conclusions of the study, and emphasize some limitations and future directions.

Related work

In this section, we review work on various methods for multiple/single document visualization, and natural language processing (NLP) techniques, such as named entity recognition, comparative sentence classification, and extractive text summarization. For visualization methods, since our proposed visualization system follows a hierarchical structure, we also take a top-down approach in reviewing existing research. That is, we will first review research that visualizes an entire topic space (composed of a collection of documents) over time, and then review research on single-document visualization at various level (topic-level, sentence-level and word-level). Research on argumentation visualization is also introduced. The storyline visualizations, from which we borrow the metaphor, are also reviewed. For NLP techniques, we will review named entity recognition, comparative sentence classification, and extractive text summarization respectively.

Multi-document visualization

Multi-document visualization in the context of scientific documents, or scientific-field evolution visualization is directly relevant to our study. Approaches in this direction involve

partitioning co-concept and/or co-keyword network into communities (scientific fields), and then investigating temporal behavior of communities, such as splitting, merging and shifting patterns of scientific fields over time (Chavalarias and Cointet 2013), and interactions between academic push and technological pull for theories (Callon et al. 1991).

To our best knowledge, the basic visual elements of scientific-field evolution visualization are usually communities. While this design clearly highlights evolving patterns of communities over time, what each community means is usually ambiguous. Moreover, entity relationships within one community and between two communities are unknown in the current visualization framework. However, entity relationships are crucial for understanding the semantics of a community, a document and a document collection. In our study, we use both community and entity as visual elements in our multi-document visualization. Their relationships are also highlighted in nested structure. This way, we visualize a document collection at meso-level that embraces both temporal patterns of communities of scientific papers, and individual entities in communities embedded in each individual scientific paper.

Single-document visualization

In this section, we review previous work on single-document visualization. Depending on the granularity of the visual elements, we divided the work into topic-level document visualization, sentence-level document visualization, and word-level visualization.

Topic-level document visualization

To facilitate exploration of a document, some applications focus on finding the latent topics of a document first, and use topics as an intermediary between words and the document for visualization. Varifocal-Reader (Koch et al. 2014) uses text-segmentation method to segment full-text into topical segments, and annotates entities such as person and location. Serendip (Alexander et al. 2014) uses statistical topic models as a bridge between words, topics and documents, and visualizes the interactions of the three elements by matrices. The advantage of this approach is that it helps to capture the topical structure of a document for easier digestion. However, it might also suffer from loss of finer-level details, such as logical chains and arguments progressively developed in sentences and entities.

Sentence-level document visualization

Some other applications focus on organizing visual elements and visualizing a document on sentence-level. One application chooses not to display all sentences plainly, but rather to display sentences using a fish-eye view so that salient passages will be highlighted as focal, and the rest will be blurred as background (Correll et al. 2011). Another application extracts quote sentences from news narratives and supports searching of quotes by speakers (Schneider et al. 2010). The strength of this approach is that finer-level details (sentences) is preserved, organized and shown to users. However, existing work makes little use of the relationships between entities and sentences, which is important for understanding the arguments of a document.

Word-level document visualization

There are multiple applications on word-level document visualization. One application finds frequent word usage patterns and highlights them in full-texts (Don et al. 2007).

Another application supports to visualize all neighboring words of a given word query in a word cloud view, and visualizes the word frequency distributions over the narrative scope (Clement et al. 2009). Another work proposes to visualize words in a document as word cloud, known as "Wordle" (Viegas et al. 2009). One work, specially tailored for play script, visualize characters-scenes as a matrix, with character on-stage-scene as highlights (Wilhelm et al. 2013). There are also some works on phonetic-levels, often tailored for poem analysis (Abdul-Rahman et al. 2013; McCurdy et al. 2016).

The advantage of visualization on word-level is that it preserves the finest-level of details. However, work mentioned above cannot fully satisfy the complex demands of scientific paper exploration. Special corpus such as poem, play, news, Bible, is usually well structured, with abundant metadata such as characters, speakers, person and location names, and set of heuristics enablesso on. However, such information is not easily available in full-text research papers without a pipeline of natural language processing. A more general framework is needed, to accommodate full-text research papers from arbitrary domains, and visualizes these papers in consistent visualizations. Moreover, most of these works do not consider the relationships between words or entities, which is crucial for understanding the document. Without highlighting the relations, users can easily get lost in separated floating entities, and therefore could not understand the entire document.

In our study, we propose single-document storylines that can draw storyline for an arbitrary research paper, not confined to any specific domain. All information the system needs is the full-text of a research paper. In our storylines, we not only visualize entities, but also highlight relationships between entities over the entire narrative. This way, the progressive development of entity relationships can be traced and understood.

Argumentation visualization

Our work may also be related to research in argumentation visualization. Research in this area attempts to visualize the structures of argumentations, usually in an interactive collaborative learning environment, to support decision making (Kirschner et al. 2012). Our work instead attempts to visualize the structures of concepts in a scientific paper via automatic natural language processing of the full text and interactive visualizations.

Storyline visualization

In our study, we borrow metaphor of "screen play" and use it in our visualization of multi-document and single-document storylines. There are already some works in storyline visualization. One work proposes to visualize the storyline of a screen play by arranging characters and scenes horizontally over time (Tanahashi and Ma 2012). More specifically, each character is represented by a curved line horizontally flowing through scenes, and each scene is represented by a rectangle bundling all character lines of this scene within it. Another work improves the previous one by optimizing multiple objective functions to make the storyline more compact and visually-pleasing (Liu et al. 2013). In our study, for multi-document storylines, we propose a new layout design inspired by the parallel coordinates (Inselberg and Dimsdale 1987) and a set of layout heuristics. For single-document storylines, we utilize a similar design as the XKCD style narrative-chart template (Bostock 2017), which uses curved-lines to represent characters, and line-bundling rectangles to represent scenes.

Named entity recognition

One objective of our system is to recognize named entities in arbitrary research papers on the fly. This is difficult since most named entity recognition methods are designed for entities of only a few domains. Majority of existing work on named entity recognition are supervised methods, including Hidden Markov Models (HMM) (Bikel et al. 1997), Maximum Entropy Models (ME) (Borthwick and Grishman 1999), Conditional Random Fields (CRF) (McCallum and Li 2003), and so on. These methods usually require a considerable amount of human-labeled data and the data is usually confined to a specific domain. However, our system needs a named entity recognizer that is nearly universal and can identify entities from arbitrary domain on the fly. In the present paper, we take advantage of the Microsoft Entity Linking Intelligent Service (ELIS),[1] which not only recognizes named entities from a wide range of topics based on Wikipedia coverage, but also links different mentions of a unique entity together.

Comparative sentence classification

Bing Liu has been one of the pioneer researchers on the topic of comparative sentence classification. In one paper, he proposes to use manual keyword list to extract candidate comparative sequences, and use frequent sequence mining to extract frequent comparative sequences, and then uses these sequences as features to train a binary classifier on labeled dataset (Jindal and Liu 2006a). In a follow-up work, he further proposes to not only classify sentences into comparatives/non-comparatives, but also extract the subjects as well as comparative relations from comparative sentences (Jindal and Liu 2006b). In the present paper, we implemented the full pipeline for comparative sentence classification as described in the paper (Jindal and Liu 2006a).

Extractive text summarization

Early work in extractive text summarization proposes to create a graph of all sentences in a document, and then select summarization sentences by random walk on the network (Erkan and Radev 2004; Mihalcea and Tarau 2004). The edges between nodes (sentences) in the network are based on semantic similarity or content overlap between two sentences. TextRank measures sentence similarity based on word co-occurrences (Mihalcea and Tarau 2004). LexRank uses cosine similarity of TF-IDF vectors for each pair of sentences (Erkan and Radev 2004). The graph-based text summarization is designed to mainly maximize the coverage of summarization sentences in original text.

Later work introduces another metric, namely diversity, to reduce redundancy in summary, especially for multi-document summarization. A method called Maximum Marginal Relevance (MMR) was proposed to incorporate the diversity in summarization (Carbonell and Goldstein 1998). This method uses a weighted combination of both candidate's similarity to a query (relevancy) and its similarity to the summary as of now (diversity) to rank candidate sentences (Carbonell and Goldstein 1998). SumBasic uses frequency alone as powerful feature in summary creation that both satisfies coverage and diversity (Nenkova and Vanderwende 2005). GRASSHOPPER further incorporates coverage and diversity in a unified framework using absorbing Markov chain random walks (Zhu et al. 2007).

[1] https://azure.microsoft.com/en-us/services/cognitive-services/entity-linking-intelligence-service/.

The state of the art methods for multi-document summarization are mixtures of submodular functions. If the objective function of summarization is a monotone submodular function, then a greedy algorithm can find an approximate solution guaranteed to be very close to the global optimal (Nemhauser et al. 1978). Researchers have proved that many existing summarization systems can be considered as instances of submodular functions (Lin and Bilmes 2011).

In our study, we modified the SumBasic algorithm to cover as many major named entities as possible, instead of plain words in sentences. We choose SumBasic because it is both simple and effective, and more importantly it can be easily adapted to entity-based algorithm which fits the needs of our system.

System pipeline

In this section, we describe the pipeline of the LitStoryTeller+ system, through which full-text data input is transformed to multiple-level storyline visualizations.

The pipeline of the LitStoryTeller+ is depicted in Fig. 1. The workflow goes from left to right, starting from uploading one or more scientific paper full-texts into the system (blue rectangles). The uploaded full-texts are first pre-processed into normalized texts, and then fed into four components. The first component, namely entity-linking component, identifies and links mentions of each named entity in the full-texts to their corresponding entity. The second component, community detection component, performs network partition on the co-occurrence network of entities and assigns a community label to each entity in the network. The third component generates extractive summarizations from a collection of full-texts, based on the major entities identified in first component with adapted SumBasic algorithm (Nenkova and Vanderwende 2005). The fourth component classifies each sentence in the full-text into comparative/non-comparative categories, based on a binary classifier with frequent sequence patterns as features.

Based on the output of the four components, three types of visualizations are generated. The first visualization, multi-document co-word (entity) network, displays a co-occurrence network of all entities in the full-texts, and colors each entity based on its community. This visualization helps the readers to have an overall impression of what are the major cliques of entities mentioned in this document collection. The second visualization, multi-document storylines, displays the storyline of entities across publications at different time points. This view is also accompanied with an extractive summarizer, which extracts N sentences from each full-text paper, and displays the summary as a ranked sentence list. This view enables users to trace the entity associations across multiple papers in a document collection, by reading the storylines over time, in the context of corresponding summaries. The third visualization displays the storyline of entities in a single paper in its narrative order, at paragraph and sentence granularities. This view enables user to drill down to details of the interactive relationships between two or more entities, in its original narrative context.

System components

In this section, we describe the details of each component in the pipeline of LitStoryTeller+. The pipeline includes pre-processing component, entity-linking component, community detection component, entity-based multi-document summarization component

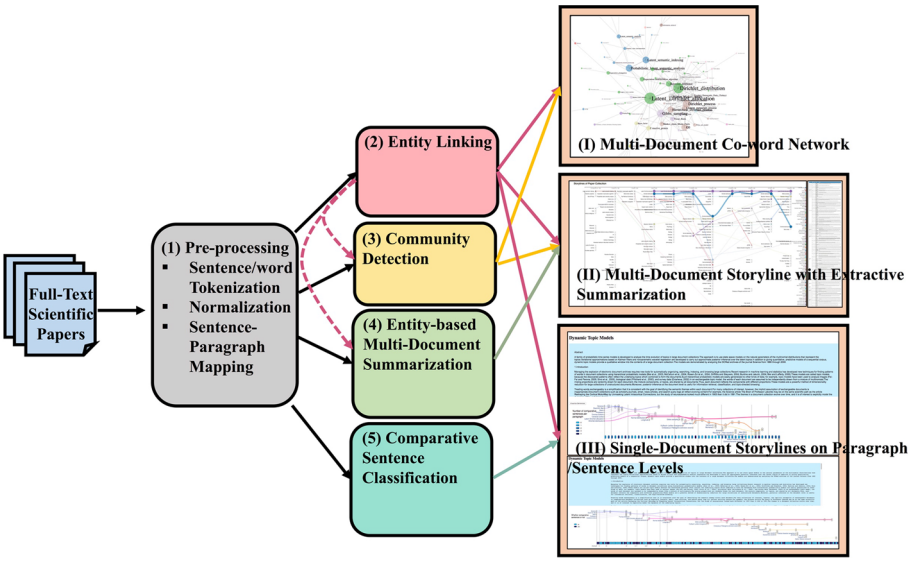

Fig. 1 Pipeline of the LitStoryTeller+ system (from full-text to storyline visualizations). (Color figure online)

and comparative sentence classification component. The final visualizations include multi-document co-word network, multi-document storylines and single-document storylines at multiple granularities.

Full-text pre-processing

The aim of the pre-processing component is to transform the full-text into multiple normalized and structured formats that are fed into different components at later stage. First, full-texts are segmented into paragraphs using line breaks, and paragraphs into sentences using sentence tokenizers. Second, special and irregular characters are removed, and tokens are stemmed. Third, sentences are POS tagged and stored as (word, tag) tuples. Last, sentences and paragraphs are linked into nested structures for later stage.

Entity-linking

We refer to entities as important scientific concepts or terminologies discussed in a research paper. A crucial step in our pipeline is to perform entity-linking, which identifies named entities and links mentions of a named entity in full-text to a unique named entity. This step recognizes a list of named entities along with their offsets in original text, which serve as inputs for downstream tasks, such as community detection task, temporal story-telling task, and so on.

To recognize named entities in scientific papers from arbitrary domains on the fly is a challenging task, since most named entity recognizers are confined to limited domains. In the present study, we take advantage of the Microsoft Entity Linking Intelligent Service

[2] https://azure.microsoft.com/en-us/services/cognitive-services/entity-linking-intelligence-service/.

(ELIS),[2] which can recognize and identify each separate entity based on its context (surrounding text).

More specifically, we concatenate a collection of documents in up to 10,000 characters to a batch, and send it to the ELIS API, to identify named entities. Before sending the batch, we also record the word-sentence-paragraph nested mapping structures. When mentions of named entities are returned, together with their unique entity identifier and offsets in the text, we can then locate the named entities in each sentence and each paragraph. This enables us to collect co-occurrences of entities at both sentence-level and paragraph-level, which serve as input for downstream tasks.

Community detection

Communities in a network are composed of nodes joined together in tightly knit groups, between which there are only looser connections (Girvan and Newman 2002). In our study, a community represents a set of entities tightly associated in the co-occurrence network, thus representing a topic with associated entities.

Given the entity co-occurrence network on sentence-level and paragraph-level built from the entity-linking component, we utilize the Louvain algorithm (Blondel et al. 2008) to detect communities in an entity-co-occurrence network. The detected communities are then used to color the nodes in a multi-document co-word network (visualization-I in Fig. 1), and group nodes in each column in a multi-document storyline view (visualization-II in Fig. 1).

Entity-based multi-document summarization

The objective of this component is to automatically generate a summary of a document collection, by extracting limited numbers of sentences from each document. The extracted summary should not only cover the major entities of each paper, but also emphasize important topics that are consistently discussed across all documents in a collection.

To generate such summary, we calculate a score for each sentence in original document based on multiple metrics, and then use an entity-based SumBasic algorithm (Nenkova and Vanderwende 2005) to select sentence for summarization. The metrics used for scoring a sentence include how many important entities the sentence contains (entity-coverage), whether the sentence is comparative (comparativeness) and the original rank of the sentence in the document.

More specifically, for a document collection $\mathcal{D} = \{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_\mathrm{n}\}$, where each document $\mathrm{d_i}$ has a set of sentences $\mathcal{S} = \{\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_{\mathrm{d_i}}\}$, and each sentence contains a set of entities $\mathrm{E}(\mathrm{d_i}) = \{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_{|\mathrm{s_i}|}\}$, the entity-based SumBasic algorithm includes the following steps:

1. **Initial entity weighting** For all entities in the document collection $\mathcal{D}$, we derive a weight for each entity based on the number of documents containing this entity: $\omega(e_j) = |\mathcal{D}_{e_j}|$, $|\mathcal{D}_{e_j} = \{d_k | e_j \in d_k\}|$;
2. **Sentence scoring** For each sentence s_k in each document d_i, we calculate an entity-coverage score $y_{entity-coverage}(s_k) = \sum_{e_j \in s_k} \omega(e_j)$, which sums up the weights of entity contained in this sentence, and a comparative-sentence score $y_{comparativeness}(s_k) = f(x) = \begin{cases} 1, s_k is comparative \\ 0, otherwise \end{cases}$, depending on the output from the

comparative-sentence classification component, and add the two scores up to obtain an overall sentence score $y(s_k) = y_{entity-coverage}(s_k) + y_{comparativeness}(s_k)$;

3. **Sentence selection** For each document, we rank all sentences based on the overall sentence score $y(s_k)$ in descending order. If multiple sentences have identical scores, they are sorted in their original rank in original text. Then the first sentence is selected into our summary;
4. **Entity re-weighting** Since the selected sentence already covers some large-weight entities, the weights of these entities should be decreased to avoid "redundancy" in summary. To achieve this, the weights of entities that are already selected are down-sampled: $\omega^{t+1}(e_j) = 0.85 \cdot \omega^t(e_j)$. The value 0.85 is empirically set and can be tuned based on specific dataset through interactions with user;
5. Go back to step (2) until a desired number of sentences are selected for each document.

The final summarization consistent of N sentences for each document, with its corresponding overall scores. This summarization provides context for the visualization-II: multi-document storyline view, which will be detailed later.

Comparative sentence classification

In some research papers, comparative sentences usually convey important information. For example, in sentences such as "the association between X and Y is stronger than that of X and Z", and "our model M scientifically outperforms baselines A, B and C", the comparative sentences indicate important findings or conclusions of a research paper. Therefore, in our study, we also perform comparative-sentence classification for each sentence in each document. The predicted comparativeness can be then used in the entity-based multi-document summarization component mentioned above, and be used in the single-document storyline view, which will be discuss in detail later.

To achieve this goal, we first train a comparative-sentence classifier based on a labeled corpus, and then use the trained classifier to predict sentence comparativeness on the fly. More specifically, given a training corpus of set of sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$, each with a label of comparativeness $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}, c_i \in \{0, 1\}$, we perform the following tasks:

1. **Constructing comparative keyword-list** A list of keywords indicating a comparative relationship should be constructed manually as the starting point. In the original work (Jindal and Liu 2006a) to identify comparative sentences, three categories of keywords are collected, namely adjectival/adverbial comparatives, single-verb keywords, and phrase-keywords. Besides these keywords, we added four keywords: "fail", "gain", "over" and "contrast" based on our observations of comparative sentences in research papers.
2. **Extracting candidate comparative sequences** Given a keyword-list $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$, and a set of sentence that are POS tagged, we scan each sentence to see if it contains any keyword, and if so, we extract a sequence of this keyword containing POS tags. For example, for the following sentence, "The concatenated features *outperform* the original features", with POS tag sequence (DT)(JJ)(NNS)(VBP)(DT)(JJ)(NNS), we extract a sequence [(DT), (JJ), (NNS), ('outperform', VBP), (DT), (JJ), (NNS)], where the window size is 3, and the central keyword is "outperform".

3. **Frequent comparative sequence pattern mining** The candidate sequences extracted above are not always typical patterns of comparative sentences. We need to mind the frequent patterns on the candidate sequences to get the most typical patterns. We adopt the PrefixSpan algorithm (Han et al. 2001) to mine frequent sequence patterns from all candidate sequences generated in the last step. The PrefixSpan algorithm utilizes projection of search space into prefix sequences to reduce the number of candidate subsequence generations (Han et al. 2001). In our PrefixsPan implementation, we set the minimum support for frequent sequence set to be 0.1, and the minimum confidence set to be 0.6. The outcome of this step is a set of keyword-POS tag sequences $\mathcal{P} = \{p_1, p_2, \ldots, p_{|\mathcal{P}|}\}$, where each frequent sequence pattern $p_i = [\ldots, pos_{i-2}, pos_{i-1}, keyword_i, pos_{i+1}, pos_{i+2}, \ldots]$ is most likely to be a sequence pattern of comparative sentence.
4. **Feature engineering** Given the frequent sequences $\mathcal{P} = \{p_1, p_2, \ldots, p_{|\mathcal{P}|}\}$ from last step, we treat each frequent sequence p_i as a unique feature for our classifier. In other words, the feature vector in our classifier is a vector of values (0 or 1) indicating whether a sentence satisfies any frequent sequences. For example, if a sentence s_k satisfies frequent sequence p_1, p_3 and p_5, then the feature vector x_{s_k} becomes $[1, 0, 1, 0, 1]$, assuming we only have 5 frequent sequences in $\mathcal{P}$, and p_2 and p_4 are not satisfied.
5. **Training classifier** Given a set of feature vectors $\mathcal{X} = \{x_{s_1}, x_{s_2}, \ldots, x_{s_n}\}$ for a set of sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$, and a set of corresponding labels $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}, c_i \in \{0, 1\}$, we train a Bayes Classifier based on given labels similar to previous work (Jindal and Liu 2006a). SVM and Logistic Regression classifier are also experimented with inferior performances compared to Bayes Classifier. We manually labelled 286 sentences from research papers, and feed the labeled sentences into our classifier. The accuracy of 5-fold cross-validation is (0.84 ± 0.02) for the Bayes classifier.
6. **New sentence prediction** In the prediction stage, each sentence is first POS-tagged and stemmed. Then the sentence is transformed to a feature vector by checking which frequent sequence patterns this sentence satisfies. Then the feature vector is fed into the trained Bayes classifier to generate a prediction, namely comparative (1) or non-comparative (0).

This comparativeness value is used in the multi-document summarization component, namely if a sentence not only covers major entities, but also is comparative, such sentence will be given higher priority in final summary. This comparative value is also used in the single-document storyline view, which will be detailed later.

Visualization I: Multi-document co-word network

As in Fig. 2, this visualization draws an entity co-occurrences network based on full-text of a document collection. Each node represents an entity, and each link represents a co-occurrence relationship between two entities. The size of an entity represents its frequency in the document collection, and the color of an entity represents its community. The thickness of a link represents the number of co-occurrences of two entities. The layout of the network is calculated with a force-layout algorithm (Bostock 2016) to minimize the number of crossings of edges in a network by optimizing energy functions (Kobourov 2012).

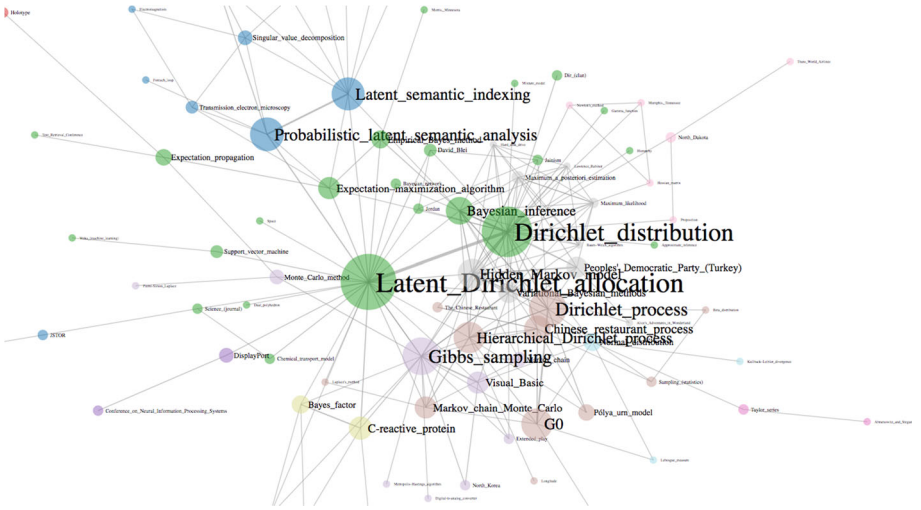

Fig. 2 Visualization-I: multi-document co-word network. (Color figure online)

This visualization is designed to offer a general impression of all major topics (in the form of entity communities) discussed in the document collection. Moreover, it emphasizes entities who occur a lot (large nodes), who connects with others a lot (nodes with many outgoing links), and who fills the gap (nodes connecting two communities) as pivot point (Chen 2004).

Nevertheless, this visualization merges entity co-occurrences of different scientific papers at different time points into one network, thus the temporal information is omitted. To overcome this, we propose the visualization II: multi-document storylines with extractive summarization.

Visualization II: Multi-document storyline with extractive summarization

This visualization consists of two parts, namely the multi-document storyline view, and the extractive summarization view.

Multi-document storyline

This visualization is designed to tell a story about entities and their temporal associations in a document collection. We borrow the "storyline" metaphor from theatre play. As in traditional "storyline" of a novel or a play, there are characters and scenes, and there are beginning, development, turning point, climax and conclusion in a plot. An evolving research topic (a collection of scientific papers) shares similarity with the formality of a play. Essentially, a plot (research topic) evolves around its main characters (key entities) at different scenes (papers), where characters have dynamic interactions with each other. As the story plot goes, more characters (entities) may come into play and interact with existing entities, and some may take a bow and leave due to lack of further investigation.

Following this analogous metaphor, we design our storyline visualization for a document collection as in Fig. 3. The visualization is interpreted as follows. First, from left to right, each column represents a scene (research paper), with the (author, publication year) tagged

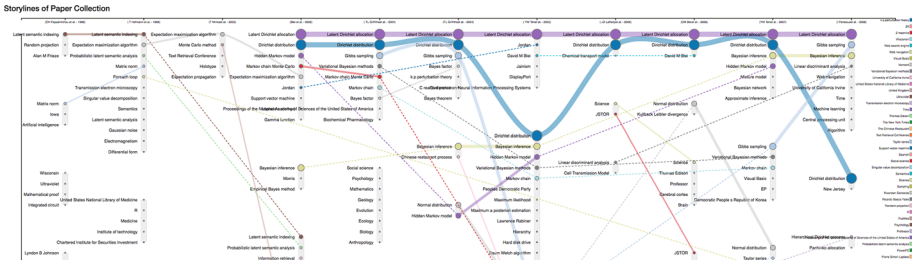

Fig. 3 Multi-document storyline view. (Color figure online)

on top of each column. Second, for each column, there are one or more long rectangles vertically aligned from top to bottom, each with one or more circles in it. Each rectangle represents a group of characters (community) of this scene (research paper), and each circle represents a character (entity) in this group (community). The size of the circle represents the frequency of the entity in the document collection, and the color of the circles is used to distinguish each entity. Third, a solid curved line connects circles of one same character (entity) at different scenes (papers), if the columns are consecutive. Otherwise, a dashed straight line is used instead. The line is the storyline of this character (entity) over time. The color of the line is the same as the color of the circle for each entity. The thickness of the line represents the frequency of the entity.To best display the storylines of entities at different time points, highlight important storylines, and minimize line-crossings, we use several heuristics for the layout of the community rectangles and entity circles:

1. **Entity layout** For entity circles in a community rectangle, we arrange their order from top to bottom by frequencies of entities in descending order. In other words, entities that frequently appear in the document collection are placed on top of the community as highlights and vice visa. If multiple entities have the same frequency, we will arrange them in alphabetical order.
2. **Community layout** For community rectangles in a single-paper column, we arrange their order from top to bottom by the frequency of its first entity. If several communities have the same frequency for their first entity, we will arrange them by their first entity name in alphabetical order.

This set of heuristics enables entities that are frequently discussed in multiple papers to be placed on top and highlighted as major storylines. In the meantime, associations of these entities at different time points are indicated by communities. This visualization enables user to trace the major entity storylines and the associated entities at different time points.

Although this visualization tells a vivid story of major entities in terms of their temporal associations in a document collection, users may need more information about what these interactions mean, to fully understand the story. To achieve this goal, we provide the complementary extractive summarization view as the context for reading the multi-document storylines.

Multi-document extractive summarization

As in Fig. 4, we provide a summarization view as the supporting context for the multi-document storyline view. The information of this view is from the output of the entity-based multi-document summarization component. For each single paper, a limited set of sentences are displayed in descending order of their scores as summary. The score,

Article	Sentence id	Score	Entities	Sentence
(CH Papadimitriou et al. - 1998)	68	7.25	Integrated circuit(1), Iowa(1), Artificial intelligence(1), Matrix norm(2), Comparative:1	The following well-known theorem gives us some idea (the subscript F denotes the Frobenius norm), TIrcorom 1 (l%kaTt and Young, see <15/) Among all n x m matrices 0 o rank at most k, Ak is the one that minimizes /iA - C((, = Ci,, (Ai,j - Ci,J)2- Therefore, LSI preserves (to the extent possible) the relative distances (and hence, presumably, the retrieval capabilities) in the term-document matrix while projecting it to a lower-dimensional space.
	3	7	Latent semantic indexing(3), Comparative:1	We also propose the technique of random projection as a way of speeding up LSI.
	11	5	Information retrieval(2), Comparative:1	However, the field of information retrieval has been evolving in directions that bring it closer to databases.

Fig. 4 Multi-document extractive summarization view

contained entities of each sentence, and comparativeness are also displayed in this view. Using the multi-document storyline view and this summarization view together, users can have a better understanding of the stories about this document collection.

One thing missing from the visualization II is how two or more entities are associated in their original paper with broader context. In other words, what did the paper say before and after the sentences that associate these entities? This broader context may be important to understand how these entities come to their associations, following a narrative driven by logical thinking process. This can only be understood by examining the original full-text of a single paper. Therefore, we propose the visualization III: single-paper storyline view.

Visualization III: Single-document storyline on paragraph/sentence-levels

This visualization consists of two parts, namely the single-document storyline view, and the single-document text view.

Single-document storyline view

For single-document storyline design, we also use the "storyline" metaphor. That is, to articulate a research idea in the narrative of a single paper, an author needs to demonstrate key entities and their associations progressively throughout the abstract, introduction, related work, methodology, experiment, discussion, conclusion and references of the research paper. These sections can be regarded as grand scenes at a coarse-level. Further, each paragraph and each sentence can be considered as scenes at fine-level.

Following this design intuition, we propose single-document storyline view at paragraph and sentence-level, as in Fig. 5. More specifically, the visualization is interpreted as follows. As in Fig. 5, first, from left to right, each vertical grey line represents a grand scene (section), with sub heading tagged on the bottom of each line of section. Second, from left to right, each small blue rectangle of different shades at the bottom of the view represents a scene (paragraph). When hovered over, a tooltip displays the full-text of this paragraph. Third, from left to right, if a scene has at least two characters (entities) in it, it is considered a major scene, and a transparent rectangle will be displayed and vertically aligned with the corresponding blue rectangle at the bottom. Within the major scene rectangle, there are a set of points vertically aligned, which represents characters (entities) co-occurring in this scene (paragraph). Fourth, a solid curved line connects points of one same character (entity) at different scenes (paragraphs). The line is the storyline of this character (entity) throughout the discourse in the narrative. The color of the line is used to distinguish an entity. The thickness of the line represents the frequency of the entity. Last, the shade of scene (blue rectangle) at the bottom represents comparativeness of the paragraph, by aggregating the number of comparative sentences in the paragraph. We also

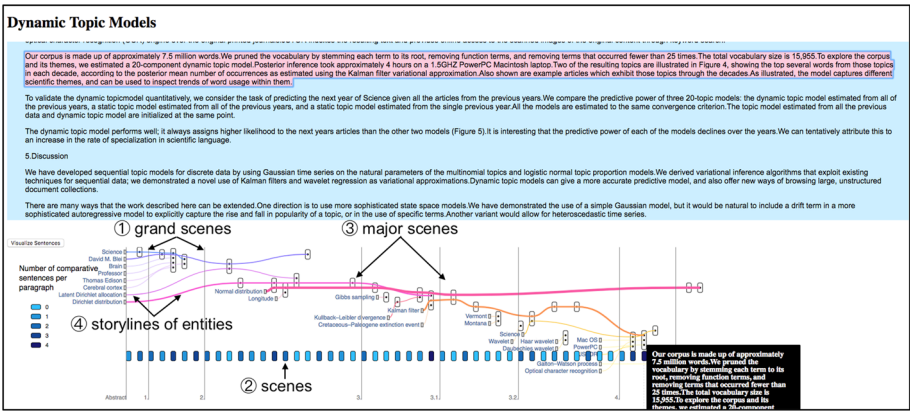

Fig. 5 Single-document storyline view at paragraph/sentence-level. (Color figure online)

visualize storylines at sentence-level. Most elements are identical to the storylines at paragraph-level, and the only difference is that instead of representing paragraphs, the blue rectangles at the bottom now represent sentences. Therefore, the shade of the blue rectangle represents whether it is a comparative sentence, instead of an aggregated comparativeness.

Although this visualization tells us a story of entities in terms of their temporal associations in a single research paper, users may need more information about what these interactions mean, to fully understand the story. To achieve this goal, we provide the complementary single-document view as the context for reading the single-document storylines.

Single-document text view

As in Fig. 5, to give context for our single-document storylines, we provide a single-document text view. Initially, the text view displays the full-text of a research paper with current focus on its beginning. When a specific scene element, either paragraph or sentence, is clicked in the single-document storyline view, the view will automatically jump to and highlight corresponding paragraph/sentence.

This text view enables users to understand how entities are associated with each other in the scene by reading the highlighted text in the text view. Moreover, the text view helps users to get a broader context of the current scene, by reading the paragraphs/sentences before and after the current scene.

Story reading strategy

The three visualizations (I, II, III) may seem overwhelming to read at first glance. A strategy is needed to prioritize which visualizations and visual elements to be read first to best understand the visualizations. Therefore, we propose a three-step visual reading strategy, namely identifying major entities and communities with visualization-I, tracing and understanding evolving entity associations over time with visualization II, and to drill down to the context of entity associations in its original context, together with the antecedents and consequences of the association. More specifically, the strategy is as follows:

1. **Identifying major entities and communities.** When exploring a document collection for the first time, users need to know what entities are involved in this document collection, which entities are important, and what associations are formed among these entities. With the help of visualization-I, users can identify major entities depending on various network characteristics, such as nodes of larger sizes in the network, nodes with many outgoing links, and nodes connecting two or more communities. Users can also easily identify associations in the network by looking at entity of the same color, which means that they form a community and are frequently associated in full-text.
2. **Tracing and understanding evolving entity associations.** When users have kept in mind the major entities and major associations from visualization-I, the users need to know how these associations are formed over time through different milestone research papers. In other words, not all associations are built all at once, but are formed progressively over time. This can be read in visualization-II, where entity associations can be traced by tracking the storylines of corresponding entities. Moreover, to get a more concrete idea of what each association means at each time point, the extractive summarization view provides the sentences where the entities co-occur in their original text, so that users can better understand the evolving associations.
3. **Drilling down to context of entity associations in its original context.** Given the evolving traces of entity associations and their corresponding context in visualization-II, users may still need more information to fully understand how each association come into being. In other words, there is a hidden logical chain in original narrative to accumulatively build up evidences to the formation of an association. Without understanding this logical chain, users may only leave with impression of "what it is" but not "why and how it is". This information can be found in visualization-III, where users can read through the storylines of entities to know how the associations are built following a hidden logical chain. In this way, users only need to focus on paragraphs/sentences that the hidden logical chain is embodied in, and skip other information for now.

As in Fig. 6, the story-reading strategy we proposed is hierarchical in nature. Visualization-I takes a snapshot of all entities and entity associations over time into one single network, visualization-II adds one dimension of time to visualization-I and stretches the entity associations along time axis, and visualization-III further adds one dimension of narrative order and stretches entity associations at one time point along the narrative axis.

Hypothetical Walkthrough-I: Storylines of evolving topic models

In this section, we demonstrate the use of our system with a comprehensive hypothetical walkthrough. We begin by describing a use scenario of a typical user and a set of research. Then we demonstrate how to use the "top-down" story-reading strategy to answer these research questions with our system.

Use scenario

We draw profile for a hypothetical user using our system. In this scenario, a typical user would be graduate student, Emma, who is new to the research topic of topic modeling, and wants to answer several research questions in mind. More specifically, Emma has already collected a set of core papers on the "topic modeling", following a survey paper on topic modeling and advices from her advisor and senior colleagues.

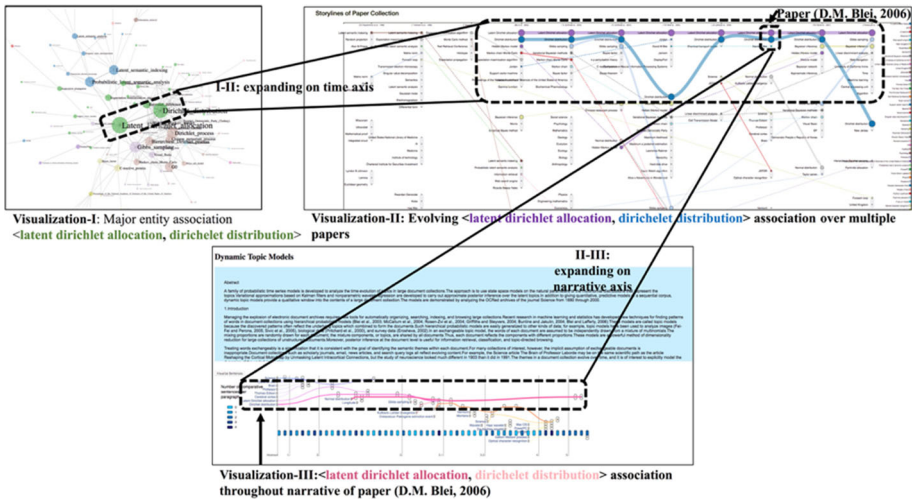

Fig. 6 Hierarchical story-reading strategy

Emma has been familiar with some classic science mapping visualization tools, such as VOSViewer (Van Eck and Waltman 2010), CiteSpace II (Chen 2006), and so on. This has enabled Emma to have a general idea of how to use science mapping tool to discover interesting patterns from scientific publications. We also suppose that Emma has a basic understanding of the subject, topic modeling, that she is ready to explore. This understanding is not in the sense of being an "expert" or "master", but a general background that is necessary for understanding the research topic of interest.

The research questions Emma has in mind are outlined as follows:

Q1: **For topic modeling, what key concepts (entities) are out there?** Are there any clusters of key concepts? Getting to know the terminologies is the first step in getting to know this new research topic. This will also help Emma to search more related scientific papers using the discovered key concepts (entities) in the future.

Q2: **What topic models are there? Can we align the various topic models temporally and progressively and have a general idea of the evolution of these models?** These research questions are important if Emma wants to have an overall historical picture of topic modeling. By aligning different topic models from the oldest ones to the newest ones, and having a general idea of how each new one is different from its old counterpart, Emma will have a deeper understanding of the topic.

Q3: **If we know the progressive alignment of these topic models, how exactly is each new model built on top of old model? Why and how the new model works better?** These research questions are crucial if Emma wants to grasp the research topic thoroughly. These questions can only be answered by getting down to details of the design of each model. By learning how why and how each new model works better than old models, Emma can not only trace the progress in this topic at its core, but also get inspiration to come up with new models on top of these existing models.

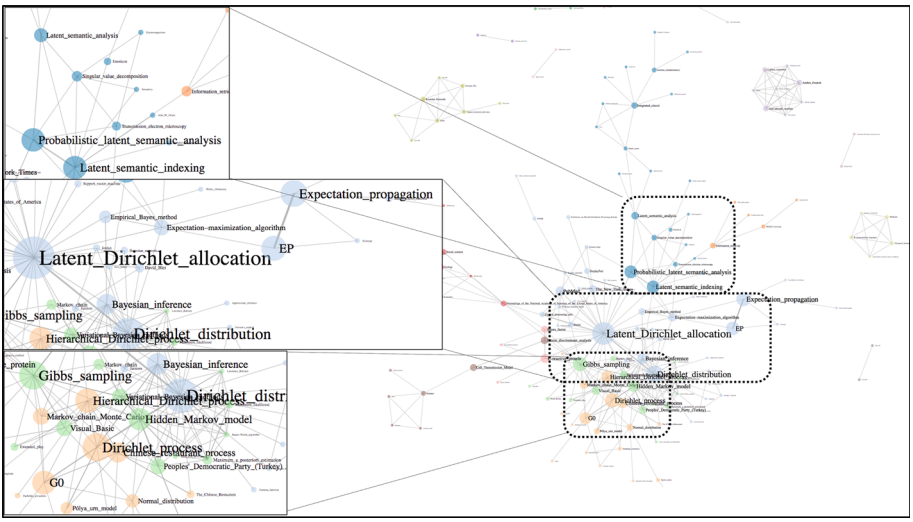

Fig. 7 Multi-document co-word network for topic modeling. (Color figure online)

Visualization-I for answering Q1

To answer the first research question, namely what are the key concepts and concept clusters in this document collection, Emma uses the visualization-I: multi-document co-word network for analysis. Emma creates visualization-I by uploading all full-texts of documents she collected, and generates the visualization with our system.

From Fig. 7, Emma immediately observes that there are several clusters of key concepts. By zooming in, Emma finds that one cluster in deep blue is about latent semantic analysis. Two of the major concepts are latent semantic indexing (LSI) and probabilistic latent semantic analysis (pLSI). Another big cluster in light blue seems to focus on dirichlet distribution, with three major concepts: latent dirichlet allocation (LDA), dirichlet distribution, and Expectation propogation (EP). The other two big clusters are tightly coupled together. The one in green is mainly about Gibbs Sampling, Hidden Markov Model, and Variantional Bayesian Inference, while the one in orange is about Dirichlet process, Chinese restaurant processs, and Hierarchical Dirichlet process (HDP).

With visualization-I, Emma has identified key concepts (big nodes in the network) and concept clusters (communities in the network) in this document collection easily. Emma takes some notes on her findings, and proceed to visualization-II.

Visualization-II for answering Q2

Recall that the second question is "What topic models are there? Can we align the various topic models temporally and progressively and have a general idea of the evolution of these models?". To answer this question, Emma uses visualization-II: multi-document storylines with extractive summarization. This visualization is generated by uploading all full-texts into the system, together with the author and publication date information manually typed in by Emma.

From Fig. 8, Emma discovers several interesting things. First, by reading the major entity storylines from left to right, Emma finds that before the year 2001, the major

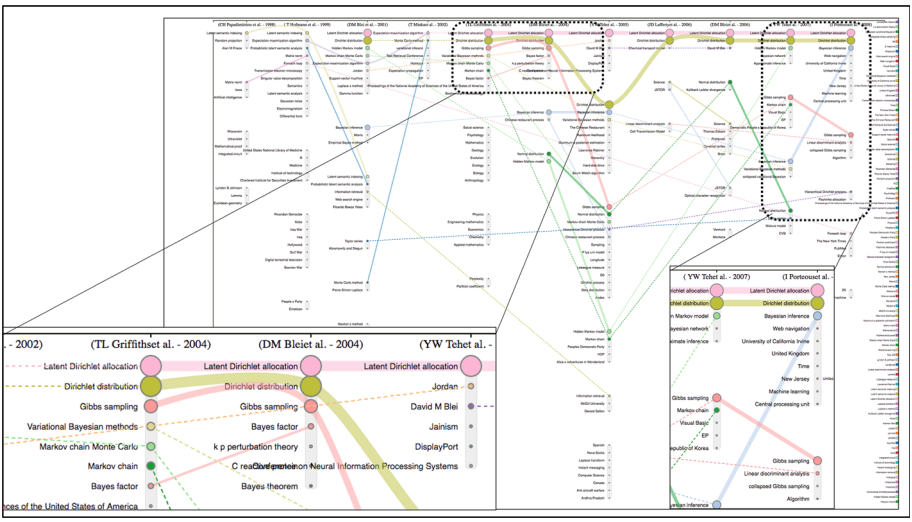

Fig. 8 Multi-document storylines for topic modeling

methods used for topic modeling are latent semantic analysis. From year 2001 and on, most of the models are seem to be about latent dirichlet allocation (LDA) and dirichlet distribution. The storylines of LDA and dirichlet distribution seems to be the dominate storylines of all entities. Between year 2004 and 2005, Gibbs Sampling enters the picture as a major storyline, and it recurs during year 2007 and 2008. Besides these storylines, Bayes inference, Hidden Markov Model, Markov Chain, Variantional Bayesian methods, normal distribution and a person name David M. Blei also have salient storylines, though may not be consecutive (dash lines across larger nodes).

Emma now have a rough idea that research on topic modeling starts with latent semantic analysis, and later transits to latent dirichlet allocation (LDA), with many variants. Gibb Sampling, and related statistical Bayesian models have a lot to do with the variants of LDA.

To further understand the transition of each year, Emma reads the Extractive Summarization view. Due to space limit, we only describe the summaries of selected papers. A full summary can be found in "Appendix". From the extractive summaries, Emma learns the following facts.

(1) For the 1999 paper (Hofmann 1999), Emma finds the following information (Table 1).

From the summary Emma gets the impression that pLSI seems to be superior to LSI in terms of solid statistical foundation, and their performances are systematically compared.

(2) For the 2004 paper (Griffiths and Steyvers 2004), where Gibb Sampling enters the picture, Emma finds the following information (Table 2).

From the summary Emma gets the impression that Gibbs Sampling seems to be an inference method for LDA, and it has been compared with two other inference algorithms: variational Bayes and expectation propagation.

(3) For the 2005 paper (Blei and Lafferty 2005) about a correlated topic model (CTM), Emma finds the following information (Table 3).

From the summary Emma gets the impression that the CTM seems to be superior to LDA based on the performances reported, and it also points out the weakness of LDA,

Table 1 Extractive summary of (Hofmann 1999)

Article	Sentence id	Score	Entities	Sentence
(Hofmann 1999)	161	8.5	Latent semantic indexing (3), Probabilistic latent semantic analysis (2), Comparative:1	The performance of PLSI has been systematically compared with the standard term matching method based on the raw term frequencies (tf) and their combination with the inverse document frequencies (tf-idf), as well as with LSI
	2	5.66	Latent semantic indexing (1.5), Singular value decomposition (1), Semantics (1), Comparative:1	In contrast to standard Latent Semantic Indexing (LSI) by Singular Value Decomposition, the probabilistic variant has a solid statistical foundation and defines a proper generative data model

Table 2 Extractive summary of (Griffiths and Steyvers 2004)

Article	Sentence id	Score	Entities	Sentence
(Griffiths and Steyvers 2004)	111	17	Variational Bayesian methods (3), Latent Dirichlet allocation (4.0), Gibbs sampling (5), Comparative:1	We applied our Gibbs sampling algorithm to this dataset, together with the two algorithms that have previously been used for inference in Latent Dirichlet Allocation: variational Bayes (1) and expectation propagation (9)

Table 3 Extractive Summary of (Blei and Lafferty 2005)

Article	Sentence id	Score	Entities	Sentence
(Blei and Lafferty 2005)	30	12	Dirichlet distribution (4.0), Latent Dirichlet allocation (4.0), Comparative:0	For the LDA model, this limitation stems from the independence assumptions implicit in the Dirichlet distribution on the topic proportions
	7	7.66	Science (2), Chemical transport model (1), Latent Dirichlet allocation (2.0), Comparative:1	The CTM gives a better fit than LDA on a collection of OCRed articles from the journal Science
	122	5.5	Linear discriminant analysis (2), Cell Transmission Model (1), Comparative:1	The CTM provides a better fit than LDA and supports more topics; the likelihood for LDA peaks near 30 topics while the likelihood for the CTM peaks close to 90 topics

Table 4 Extractive Summary of (Porteous et al. 2008)

Article	Sentence id	Score	Entities	Sentence
(Porteous et al. 2008)	239	14.66	Latent Dirichlet allocation (4.0), Linear discriminant analysis (2), Gibbs sampling (5), Comparative:0	In this paper, we have described a method for increasing the speed of LDA Gibbs sampling while providing exactly equivalent samples, thus retaining all the optimality guarantees associated with the original LDA algorithm

which is the independence assumptions on topic relatedness. Also, Emma noticed that the system has mistakenly identified "CTM" as "chemical transport model" and "cell transimission model" instead of "correlated topic models". This is probably because CTM is used more often to represent the former two concepts than the last one. Nevertheless, the system is able to use this high-frequency CTM to find important sentences as summary.

(4) For the 2008 paper (Porteous et al. 2008), where Gibbs Sampling has re-entered the picture, Emma finds the following information (Table 4).

From the summary Emma gets the impression that this paper has proposed an improved version of Gibbs Sampling for LDA, so that the speed is increased, while providing exactly equivalent samples.

At this point, Emma has a better idea about this document collection, in terms of what are the major topic models, how to align these models temporally and progressively, and how each new model is different from old models.

However, these findings have helped Emma to answer the "what" question, not exactly "why and how". In other words, for now Emma has discovered and memorized some important facts about classical topic models, but hasn't fully grasped the core principles behind these models. To understand these core principles, original text must be read, since the logical chains authors used in their writing can only be traced in the original text.

Emma takes some more notes, and proceeds to visualization-III: single-document storyline view.

Visualization-III for answering Q3

Recall that the third question is "how exactly is each new model built on top of old model? Why and how the new model works better?". To answer this question, Emma uses visualization-III: single-document storylines. This visualization is generated by uploading each individual research paper into the system, and generating the storylines for each individual paper respectively.

Emma selects the paper of probabilistic Latent Semantic Indexing (Hofmann 1999) to see how and why pLSI works better than Latent Semantic Indexing (LSI). The overview of entire single-document view is depicted in Fig. 9.

Emma starts reading the storylines at paragraph-level. As in Fig. 10, Emma finds that the storylines of pLSI and LSI have crossed at multiple paragraphs (rectangles bundling storylines), which means that pLSI and LSI have been discussed together in these paragraphs. By reading each paragraph of these crossings, and the antecedent and consecutive paragraphs of it, Emma starts to understand why pLSI works better than LSI: the objective

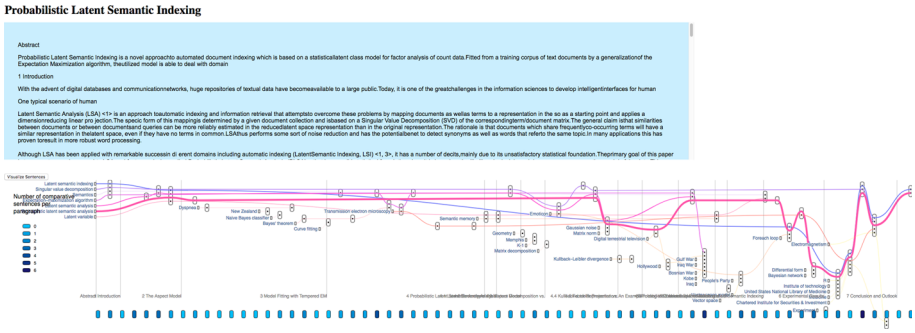

Fig. 9 Overview of single-document storyline view

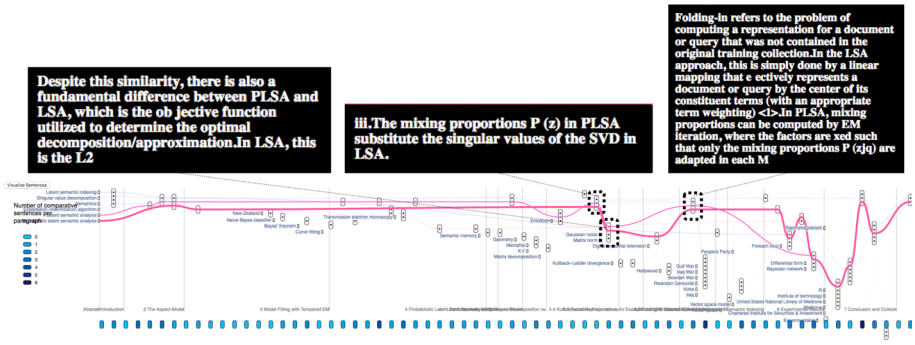

Fig. 10 Storylines of pLSI and LSI on paragraph-level

functions used in pLSI is different than that in LSI; the mixing proportions in pLSI has substituted the values in SVD for LSI; and the solution to the objective optimization is expectation maximization in pLSI, compared to SVD in LSI.

Next, Emma switches to the storylines of pLSI at sentence-level, as in Fig. 11. This time, Emma notices multiple crossings at the "Experimental results" Section. By reading the sentence of each crossing, Emma learns some "conclusive" information about the relationships between pLSI and LSI. From these reports, Emma knows that pLSI has outperformed LSI consistently in the experiments, and especially on raw term frequencies. This makes the picture complete as for how and why pLSI works better than LSI.

Emma repeats this process for every individual paper in the collection, and eventually has been able to answer all three research questions in her mind. Emma collects all the notes she took and all visualizations the system generated as a full documentation of her review process of this research topic.

Hypothetical Walkthrough-II: Historical storylines of World War II

In this section, we further demonstrate the use of our system by generating storylines from a collection of historical documents of major events in world war II. We begin by describing a hypothetical use scenario consisting of a typical user, and a set of questions to

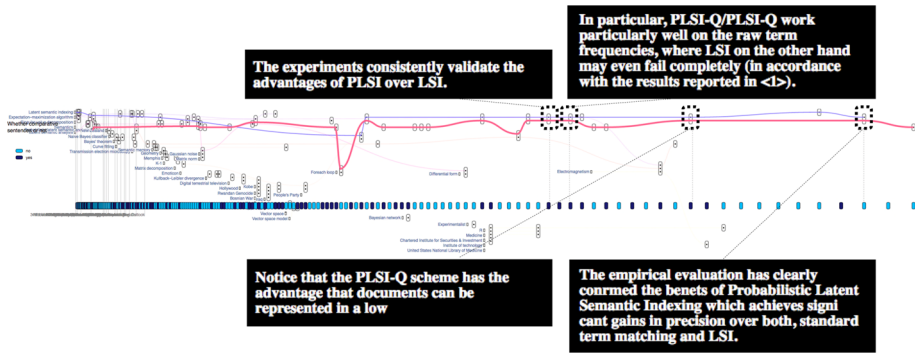

Fig. 11 Storylines of pLSI and LSI on sentence-level

be answered by our system. Then we demonstrate how to answer these questions using the "top-down" story-reading strategy with our system.

Use scenario

We draw profile for a typical user using our system. In this scenario, a typical user would be graduate student, Bob, who is already familiar with the topic of World War II (WWII), but wants to straighten up the complex relationships between events and between countries and get both an overall and detailed picture of the entire warfare. More specifically, Bob has already collected a set of documents that record major events during WWII by year from Wikipedia,[3] and wants to gain insight from this collection of documents with the help of our system.

Bob has been familiar with some classic science mapping visualization tools. This has enabled Bob to have a general idea of how to use science mapping tool to discover interesting patterns from literature. We also note that Bob has prior knowledge of the topic to be explored.

The research questions Bob has in mind are outlined as follows:

Question-1: What countries were mainly involved in World War II (WWII), and which groups of countries interacted with each other the most during the war?
Question-2: To understand the course of the warfare over time, Bob wants to consider the following sub-questions: (1) in terms of the major nations and major groups identified in visualization-I, how were they interacting with each other over time, during the war? Were there any salient collaborative or conflicting relations between nations formed over time? (2) in terms of the battle fields in WWII, did they become the major battle fields at the beginning of the warfare or at later stages? What countries were involved in each battle field? (3) how does the scale of the warfare change over time?
Question-3: Taking one-year's events for example, what are the detailed story of these events for this year? How did different nations take actions over time? How did the focus of battle field change over time?

[3] https://en.wikipedia.org/wiki/World_War_II.

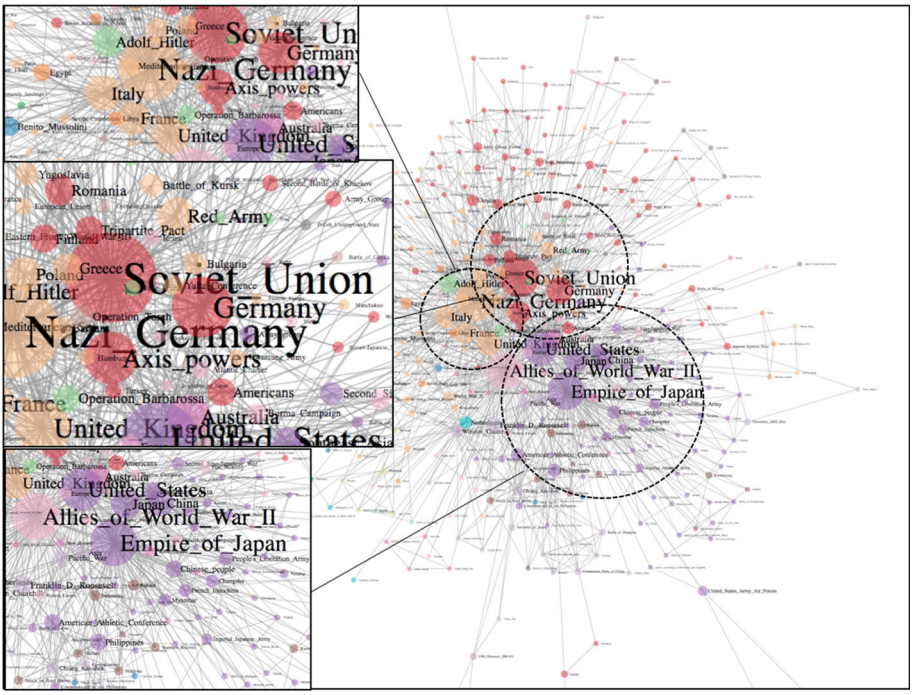

Fig. 12 Multi-document co-word network for World War II. (Color figure online)

Visualization-I for answering Q1

In order to answer the first question, namely what countries were mainly involved in world war II, and which groups of countries interacted with each other more often during the war, Bob uses visualization-I: multi-document co-word network for analysis. Bob creates this visualization by uploading all full-texts of Wikipedia documents into the system.

From Fig. 12, Bob identifies three big clusters in yellow, red and purple, of which majority are names of nations. Due to space limits, we report only bigger nodes here, although a large number of nations in smaller nodes were also impacted in the war, and not covered in full detail in the documents. The biggest nodes in the yellow cluster include Nazi Germany, United Kingdom, Italy, France, Poland, and Greece. It seems that these countries had a lot of interaction in the war. Bob thinks this is reasonable since these countries represented the western Europe battlefield, one of the major battlefield in WWII. The biggest nodes in the red cluster are Soviet Union, Germany, Romania, Finland, Americans and a military alliance: Axis Power. Bob could align these countries to the Eastern Front battlefield, where there were conflicts between European Axis powers and Soviet and other allies. The major nodes in the purple cluster are Empire of Japan (also Japan), United States, China, and Philippines. Bob thinks these countries were major countries involved in the Pacific battlefield of WWII.

With Visualization-I, Bob is able to identify the major countries involved in WWII, and the major battlefields comprised of the countries mentioned above. The next step is to look at the main course of World War II by investigating the interactive patterns between countries temporally at meso-level.

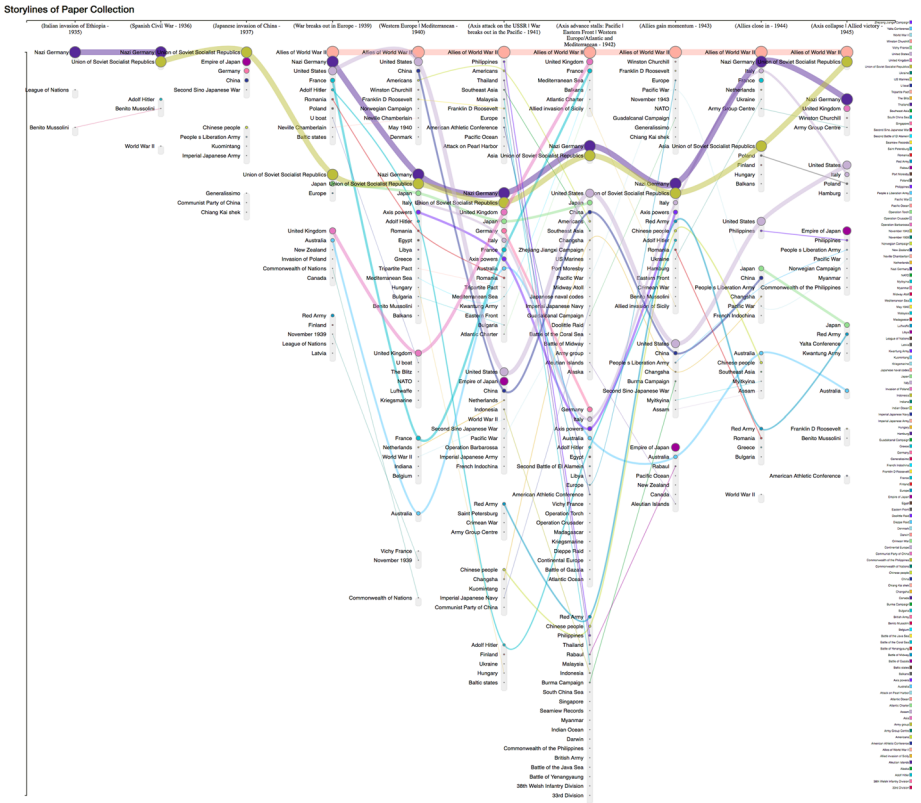

Fig. 13 Multi-document storylines for World-War-II. (Color figure online)

Visualization-II for answering Q2

From visualization-II (Fig. 13), Bob makes several observations that can offer some insights on the sub-questions of question-2.

1. **Major nation storylines** Most of the major nations (bigger circles in visualization-I) have a major storyline across time dimension, such as Nazi Germany (purple), Union of Soviet Socialist Republics-USSR (yellow green), United Kingdom (pink), United States (light purple), France (blue), Italy (grey purple), Empire of Japan (light green), China (dark blue) and so on. We can also find that the Allies were formalized around 1939, while the Axis Power came into being around 1940.

2. **Interactive patterns between nations** Based on his prior knowledge on this historical subject, Bob is also able to quickly identify two interactive patterns between nations from the visualization.

    - **Collaborative relations** Nations of such relationships are considered allies during the war, such as the allies of WWII, and the Axis Power. One example of such type of relationship can be seen between United States-China, where China's dark blue curve have almost identical trajectory with United States' light purple curve (1940-

1944). A similar pattern can be observed between Italy-Axis Power, where the grey purple curve of Italy almost has identical trajectory with the curve of Axis Power (1940-1943), although Axis Power is an alliance name. The United Kingdom-France relationship is also collaborative, with the two curves have very similar shapes across time dimension.

- **Conflicting relations** Nations of such relationships are considered enemies during the war. One example of such type of relationship between countries can be seen at Nazi Germany-USSR, where the curves of the two nations were not intertwined at first (1935-1939) but were closely joined later (1940-1943).

3. **Temporal-Geographical transitions of battle fields** From the visualization, we can also observe the temporal-geographical transitions of battle fields, from the vertically-aligned "communities" each year. For example, for Nazi Germany, it is clear that its main battle fields started in western Europe (top-most community in 1939), with nations of France, Poland, Baltic states, and so on, then expanded to Mediterranean (second community in 1940, fourth community in 1942), with nations of Egypt, Libya, Greece, and so on, and later further expanded to Eastern Front (second community in 1941, 1942 and 1943), with nations of USSR and so on. Similar patterns can be observed for Italy and Romania. For Japan, it took a different route, where it started its battle field in mainland China (first community in 1937), and expanded to Malay Peninsula (first community in 1941) and European colonies of Indonesia and French Indochina (third community in 1941), and further expanded to the Pacific, with nations of United States, Australia, and so on (third community in 1942, fourth community in 1943).

4. **Temporal change of scale of warfare** If we step back and see the overall picture, we can observe that the entire WWII started regionally at smaller scale (1939), and then ravaged the world by involving more and more nations and regions, and eventually became a global warfare at large scale (1940-1943). As major Axis Power nations surrendered, the scale of warfare shrank to be regional again and eventually died down (1944-1945).

Bob takes some notes and proceeds to read storylines of each individual year during WWII, hoping to get down to details about the course of warfare in each year.

Visualization-III for answering Q3

Bob select the year 1945, when the warfare was at its last stage, and uploads the document with title "Axis Collapse, Allied Victory" into the system. The generated storylines can be seen in Fig. 14.

From the visualization-III, Bob is able to quickly identify two main episodes of stories, namely the ending war between Nazi Germany and the Allies (including USSR and United States), which accounts for the major entanglements of storylines in the upper part of the visualization, and the ending war between Japan and the Allies (including United States, China, Philippine, USSR) which accounts for the entanglements of storylines at the lower right part of the visualization.

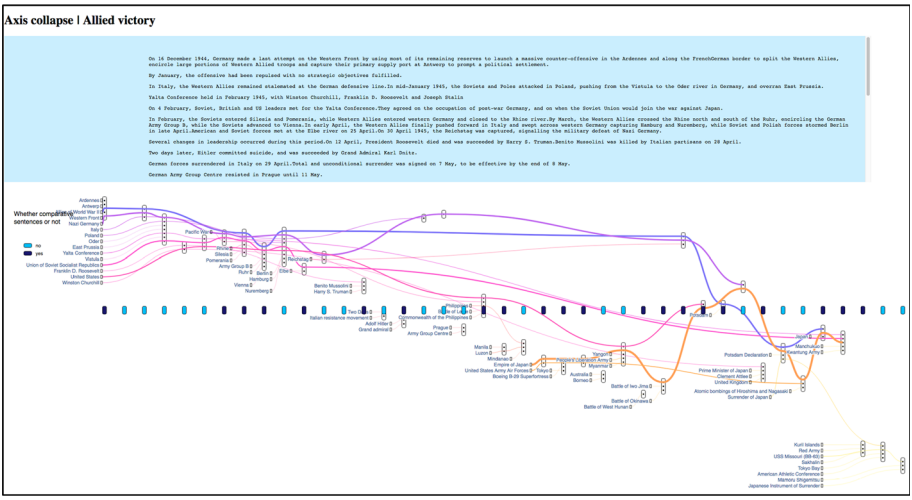

Fig. 14 Storylines of 1945: Axis Collapse and Allied Victory

Moreover, by reading over the sentences of each time points, Bob is able to track the progress of the Allies against the Axis Power. For the ending-war with Nazi Germany, as depicted in Fig. 15(1-2), the Allies worked collaboratively with USSR to fight with Nazi Germany, and eventually pushed forward in Italy and western Germany, while USSR and Poland forces stormed Berlin.

For the ending-war with Empire of Japan, as depicted in Fig. 15(3-5), United States, working collaboratively with the Philippines, China, United Kingdom and Australia, cleared Leyte and Burma from Japan, and also made air attacks to homeland Japan with United States Army Air Forces (USAAF). USSR also joined the battle field at later stage, defeated the Kwantung Army in Japanese-held Manchuria of China, after Japan rejected the call of unconditional surrender. After the USAAF atomic bombing of Hiroshima and Nagasaki, Japan surrendered, ending WWII.

At this stage, Bob understands the detailed stories of the events in year 1945, with two major storylines of ending-war between Allies and Nazi Germany and ending-war between Allies and Empire of Japan. USSR fought both in the western front against Nazi Germany with western Allies, and then fought in Manchuria against Empire of Japan later. United States fought on the Pacific battle field together with other Allies and drove Japan to final surrender.

Bob repeats this process for every individual document in the collection, and eventually has been able to answer all three questions in his mind. Bob collects all the notes he took and all visualizations the system generated as a full documentation of his review process of this topic.

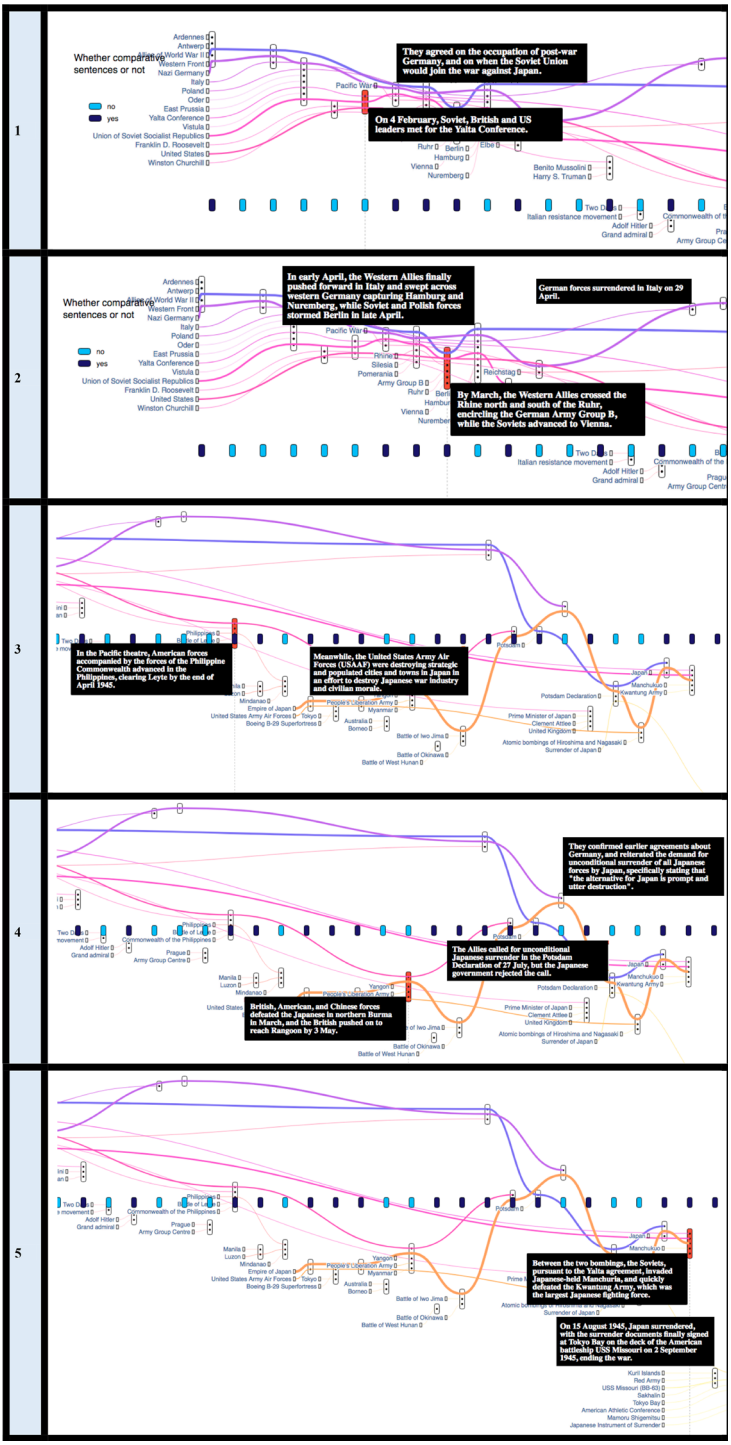

Fig. 15 Comic-style storyline reading of ending war with Nazi Germany and empire of Japan

Conclusion

In this paper, we present the LitStoryTeller+ system that can support interactive visual storytelling of scientific papers at multiple levels, by mining full-text scientific paper using several natural language processing techniques. The core idea of our system is to use entities (concepts or terminologies) and their associations in scientific papers as basic visual elements of our storylines at various levels. We believe that entities and their associations are the main embodiment of knowledge in scientific papers. Named entity recognition and community detection are performed to identify entities and their associations from arbitrary scientific papers on the fly. Moreover, text-summarization and comparative sentence classification are performed to extract rich textual information as supplementary information for visual storyline reading. A top-down story-reading strategy is also proposed to best utilize our system, starting from reading a snapshot of entity co-occurrence network for all scientific papers, to reading temporal entity storylines across multiple scientific papers at different time points, to further drilling down to single-paper storylines over the axis of narrative. Two comprehensive hypothetical walkthrough demonstrate the usefulness of our system in answering a variety of research questions, ranging from general questions like "what are the major (clusters of) entities in this paper collection?" to questions like "what are the alignment and interaction of these entities in terms of time and progress?", to detailed questions like "what are the detailed stories of entity–entity interactions at a finer-level over time?". The findings, together with the visualizations generated by our system, can be used as supplementary information for systematic review of this collection of scientific papers. However, our study also has several limitations. First, the Microsoft ELIS named entity recognizer can mis-classify or neglect some entities occasionally, especially when the entities are novel and rare. This can be partially remedied by specifying entity names by users, which is supported in our system. Second, for relationships between entities, we only use entity co-occurrences at paragraph and sentence level, instead of semantic relationships. One difficulty is that not all entities have explicit semantic relationships. In some domains, such as biomedical science, semantic relationships are prevalent and vital for understanding entity interactions. However, if we switch to a different domain such as computer science, relationships between entities are usually implicit and hard to extract. Nevertheless, we plan to explore this direction in our future work. Another future work would be identifying and visualization potential future associations between entities. Given current entity associations, potential future associations can be predicted with a variety of link prediction techniques. We plan to not only incorporate link prediction in our system, but also provide aiding visualizations to interpret the results of link prediction.

Acknowledgements This study is supported by the project "A Visual Analytic Observatory of Scientific Knowledge" funded by National Science Foundation (NSF 1633286).

Appendix

Extractive summary generated in Hypothetical Walkthrough-I

See Table 5.

Table 5 Extractive summaries (N=5)

Article	Sentence id	Score	Entities	Sentence
Papadimitriou et al. (2000)[h]	68	7.25	Integrated circuit (1), Iowa (1), Artificial intelligence (1), Matrix norm (2), Comparative:1	The following well-known theorem gives us some idea (the subscript F denotes the Frobenius norm), Tlrcorom 1 (I%kaTt and Young, see <15/) Among all n × m matrices 0 o rank at most k, Ak is the one that minimizes /iA − C ((, = Ci,, (Ai,j − Ci,J)2- Therefore, LSI preserves (to the extent possible) the relative distances (and hence, presumably, the retrieval capabilities) in the term-document matrix while projecting it to a lower-dimensional space
	3	7	Latent semantic indexing (3), Comparative:1	We also propose the technique of random projection as a way of speeding up LSI
	11	5	Information retrieval (2), Comparative:1	However, the field of information retrieval has been evolving in directions that bring it closer to databases
	21	4	Latent semantic indexing (1.5), Comparative:1	This hidden structure is not a fixed manyto-many mapping between terms and concepts, but depends critically on the corpus (document collection) in hand, and the term correlations it embodies
	185	4	Euclidean geometry (1), Lemma (1), Comparative:1	Using the above result, we can infer that with high probability, all pairwise Euclidean distances are approximately maintained under projection to a random subspace, By choosing 1 to be Q (y) in Lemma 2, with high probability the projected vectors, after scaling by a factor &ji,<=“‘ 114=“‘ 74112=“‘ i=“‘ 11%=“‘ vj<=“‘ td=“‘ style=“box-sizing: border-box;”>

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Hofmann (1999)[f]	161	8.5	Latent semantic indexing (3), Probabilistic latent semantic analysis (2), Comparative:1	The performance of PLSI has been systematically compared with the standard term matching method based on the raw term frequencies (tf) and their combination with the inverse document frequencies (tf-idf), as well as with LSI
	118	7	Bosnian War (1), Digital terrestrial television (1), Rwandan Genocide (1), Iraq War (1), Kobe (1), Comparative:1	Table 2 shows some more factors for the TDT-1 collection which clearly reect the vocabulary dealing with certain events: the war in Bosnia and Iraq, the crisis in Rwanda, and the earthquake in Kobe
	162	6	R (1), Medicine (1), Institute of technology (1), United States National Library of Medicine (1), Comparative:1	We have utilized the following four medium<=“‘ td=“‘ style=“box-sizing: border-box;”>
	2	5.66	Latent semantic indexing (1.5), Singular value decomposition (1), Semantics (1), Comparative:1	In contrast to standard Latent Semantic Indexing (LSI) by Singular Value Decomposition, the probabilistic variant has a solid statistical foundation and defines a proper generative data model
	97	5.33	Latent semantic analysis (1), Gaussian noise (1), Matrix norm (2), Comparative:0	In LSA, this is the L2<=“‘ td=“‘ style=“box-sizing: border-box;”>

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Blei et al. (2002)[c]	260	25	Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:1	In fact, by placing a Dirichlet prior on the multinomial parameter we obtain an intractable posterior in the mixture model setting, for much the same reason that one obtains an intractable posterior in the basic LDA model
	236	13.33	Empirical Bayes method (1), Latent Dirichlet allocation (4.0), Bayesian inference (5), Comparative:0	In this section we present an empirical Bayes method for parameter estimation in the LDA model (see Section 5.4 for a fuller Bayesian approach)
	279	10	Dirichlet distribution (4.0), Latent Dirichlet allocation (2.0), Comparative:1	After removing a standard list of stop words, we used the EM algorithm described in Section 5.3 to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model
	24	9.25	Latent semantic indexing (3), Bayesian inference (2.5), Comparative:1	Given a generative model of text, however, it is not clear why one should adopt the LSI methodology one can attempt to proceed more directly, fitting the model to data using maximum likelihood or Bayesian methods
	398	8.5	Latent Dirichlet allocation (1.0), Hidden Markov model (4), Comparative:1	As is the case for other mixture models, including finite mixture models and hidden Markov models, the emission probability p (wn |zn) contributes only a likelihood value to the inference procedures for LDA, and other likelihoods are readily substituted in its place

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Minka and Lafferty (2002)[g]	41	17	Dirichlet distribution (8), Comparative:1	Unfortunately, while the Dirichlet can capture variation in the p (w)s, it cannot capture co- variation, the tendency for some probabilities to move up and down together
	50	9	Dirichlet distribution (4.0), Comparative:1	The probability of a document is where the parameters are the Dirichlet parameters a and the multinomial models p (l a); denotes the (A 1)-dimensional simplex, the sample space of the Dirichlet D (l)
	80	4.5	Expectation propagation (1), Monte Carlo method (2), Comparative:0	Laplaces method using a softmax transformation, vari- ational inference, and two different Monte Carlo algorithms
	40	4	Dirichlet distribution (2.0), Comparative:0	One natural choice is the Dirichlet distribution, which is conju- gate to the multinomial
	204	3.33	Expectation propagation (0.5), Holotype (1), EP (1), Comparative:0	Figure 3 shows the test set perplexities for VB and EP; the perplexity for the EP-trained model is consistently lower than the perplexity of the VB-trained model

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Griffiths and Steyvers (2004)[e]	48	24	Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:0	In Latent Dirichlet Allocation, documents are generated by first picking a distribution over topics from a Dirichlet distribution, which determines P (z) for words in that document
	111	17	Variational Bayesian methods (3), Latent Dirichlet allocation (4.0), Gibbs sampling (5), Comparative:1	We applied our Gibbs sampling algorithm to this dataset, together with the two algorithms that have previously been used for inference in Latent Dirichlet Allocation: variational Bayes (1) and expectation propagation (9)
	199	12	Applied mathematics (1), Mathematics (1), Psychology (1), Biology (1), Geology (1), Evolution (1), Social science (1), Chemistry (1), Physics (1), Ecology (1), Comparative:1	In some cases, a single topic was the most diagnostic for several classes: topic 2, containing words relating to global climate change, was diagnostic of Ecology, Geology, and Geo- physics; topic 280, containing words relating to evolution and natural selection, was diagnostic of both Evolution and Popu- lation Biology; topic 222, containing words relating to cognitive neuroscience, was diagnostic of Psychology as both a Biological and a Social Science; topic 39, containing words relating to mathematical theory, was diagnostic of both Applied Mathe- matics and Mathematics; and topic 270, containing words having to do with spectroscopy, was diagnostic of both Chemistry and Physics

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
	135	11.5	Markov chain (3), Dirichlet distribution (4.0), Comparative:1	The statistical model we have described is conditioned on three parameters, which we have suppressed in the equations above: the Dirichlet hyperparameters and and the number of topics T. Our algorithm is easily extended to allow,, and z to be sampled, but this extension can slow the convergence of the Markov chain
	141	11	Bayesian inference (5), Comparative:1	Given values of and, the problem of choosing the appropriate value for T is a problem of model selection, which we address by using a standard method from Bayesian statistics (15)
(Griffiths and Steyvers 2004)[d]	76	24	Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:0	When the distribution p (|) is chosen to be a Dirichlet distribution, we obtain the latent Dirichlet allocation model (LDA) <11>.
	159	15.66	Bayes factor (2), Latent Dirichlet allocation (4.0), Gibbs sampling (5), Comparative:1	With the LDA model, the Bayes factors method is much slower than the CRP as it involves multiple runs of a Gibbs sampler with speed comparable to a single run of the CRP sampler.
	45	12	Normal distribution (4), Hidden Markov model (4), Comparative:0	Applications to various kinds of mixture models have begun to appear in recent years; examples include Gaussian mixture models <8>, hidden Markov models <9> and mixtures of experts <10>

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
	15	11	Bayesian inference (5), Comparative:1	We approach this model selection problem by specifying a generative probabilistic model for hierarchical structures and taking a Bayesian perspective on the problem of learning these structures from data
	80	9	Dirichlet distribution (4.0), Comparative:1	A document is generated as follows: (1) choose a path from the root of the tree to a leaf; (2) draw a vector of topic proportions from an L-dimensional Dirichlet; (3) generate the words in the document from a mixture of the topics along the path from root to leaf, with mixing proportions
Teh et al. (2005)[j]	564	27	Dirichlet distribution (8), Variational Bayesian methods (3), Maximum likelihood (1), Bayesian inference (5), Hard disk drive (1), Lawrence Rabiner (1), Maximum a posteriori estimation (1), Hidden Markov model (4), Comparative:0	Using the direct assignment sampling method for posterior predictive inference, we compared the HDD-HMM to a variety of other methods for prediction using hidden Markov models: (1) a classical HMM using maximum likelihood (ML) parameters obtained via the Baum-Welch algorithm, (2) a classical HMM using maximum a posteriori (MAP) parameters, taking the priors to be independent, symmetric Dirichlet distributions for both the transition and emission probabilities, and (3) a classical HMMtrained using an approximation to a full Bayesian analysis in particular, a variational Bayesian (VB) method

Table 5 Extractive summaries (N=5)

Article	Sentence id	Score	Entities	Sentence
	434	19	Dirichlet distribution (4.0), Latent Dirichlet allocation (8), Comparative:1	As in simpler finite mixture models, it is natural to try to extend LDA and related models by using Dirichlet processes to capture uncertainty regarding the number of mixture components
	551	13.33	Hidden Markov model (2.0), Markov chain Monte Carlo (3), Gibbs sampling (5), Comparative:0	(2002) did not present an MCMC inference algorithm for the infinite hidden Markov model, proposing instead a heuristic approximation to Gibbs sampling
	54	9	G0 (1), Lebesgue measure (1), Normal distribution (4), Comparative:1	That this simple hierarchical approach will not solve our problem can be observed by considering the case in which G0 () is absolutely continuous with respect to Lebesgue measure for almost all (e.g., G0 is Gaussian with mean)
	435	9	Latent Dirichlet allocation (4.0), Comparative:1	This is somewhat more difficult than in the case of a simple mixture model, however, because in the LDA model the documents have document-specific mixing proportions
Blei and Lafferty (2005)[a]	23	24	Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:0	The topics are shared by all documents in the collection; the topic proportions are document-specific and randomly drawn from a Dirichlet distribution
	30	12	Dirichlet distribution (4.0), Latent Dirichlet allocation (4.0), Comparative:0	For the LDA model, this limitation stems from the independence assumptions implicit in the Dirichlet distribution on the topic proportions

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
	7	7.66	Science (2), Chemical transport model (1), Latent Dirichlet allocation (2.0), Comparative:1	The CTM gives a better fit than LDA on a collection of OCRed articles from the journal Science
	80	6	Markov chain Monte Carlo (3), Comparative:0	For many problems this optimization problem is computationally manageable, while standard methods, such as Markov Chain Monte Carlo, are impractical
	122	5.5	Linear discriminant analysis (2), Cell Transmission Model (1), Comparative:1	The CTM provides a better fit than LDA and supports more topics; the likelihood for LDA peaks near 30 topics while the likelihood for the CTM peaks close to 90 topics.
Blei and Lafferty (2006)[b]	52	24	Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:0	In LDA, the document-specific topic proportions are drawn from a Dirichlet distribution.
	76	10	Gibbs sampling (5), Comparative:0	While Gibbs sampling has been effectively used for static topic models (Griffiths and Steyvers 2004), nonconjugacy makes sampling methods more difficult for this dynamic model.
	36	9	Dirichlet distribution (4.0), Comparative:1	Choose topic proportions from a distribution over the (K 1)-simplex, such as a Dirichlet.
	48	9	Normal distribution (4), Comparative:1	Instead, we chain the natural parameters of each topic t,k in a state space model that evolves with Gaussian noise; the simplest version of such a model is t,k|t1,kN (t1,k, 2I) .
	32	8	Latent Dirichlet allocation (4.0), Comparative:0	First, we review the underlying statistical assumptions of a static topic model, such as latent Dirichlet allocation (LDA)

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Teh et al. (2007)[k]	132	26.66	Dirichlet distribution (8), Latent Dirichlet allocation (8), Hidden Markov model (4), Comparative:0	Specific examples include various extensions of LDA <4, 13> hidden Markov models with dis- crete outputs, and mixed-membership models with Dirichlet distributed mixture coefficients <14>.
	6	14.5	Dirichlet distribution (4.0), Bayesian inference (5), Comparative:1	In a Bayesian setting it is convenient to endow these models with Dirichlet priors over the parameters as they are conjugate to the multinomial distributions over the discrete random variables <1>
	11	13.5	EP (1), Visual Basic (1), Variational Bayesian methods (3), Gibbs sampling (5), Comparative:1	A host of inference algorithms have been proposed, ranging from variational Bayesian (VB) inference <2>, expectation propagation (EP) <7> to collapsed Gibbs sampling <5>
	89	10	Normal distribution (4), Taylor series (2), Comparative:1	We further approximate the function log (+ nij) using a second-order Taylor expansion about, and evaluate its expectation under the Gaussian approximation
	1	9.33	Dirichlet distribution (2.0), Bayesian network (1), Latent Dirichlet allocation (4.0), Comparative:0	Latent Dirichlet allocation (LDA) is a Bayesian network that has recently gained much popularity in applications ranging from document modeling to computer vision

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
Porteous et al. (2008)[i]	47	22.66	New Jersey (1), Dirichlet distribution (8), Latent Dirichlet allocation (8), Comparative:0	The LDA model is equivalent to the following generative process for words and documents: For each of Nj words in document j 1. sample a topic zij Multinomial (j) 2. sample a word xij Multinomial (zij) where the parameters of the multinomials for topics in a document j and words in a topic k have Dirichlet priors
	239	14.66	Latent Dirichlet allocation (4.0), Linear discriminant analysis (2), Gibbs sampling (5), Comparative:0	In this paper, we have described a method for increasing the speed of LDA Gibbs sampling while providing exactly equivalent samples, thus retaining all the optimality guarantees associated with the original LDA algorithm
	14	11.5	Latent Dirichlet allocation (2.0), Bayesian inference (5), Comparative:1	Blei et al <3> introduced the LDA model within a general Bayesian framework and developed a variational algorithm for learning the model from data
	39	9	Normal distribution (4), Comparative:1	In <8>, a similar branch-and-bound method is used to compute approximate probabilities and draw approximate samples from the product of several Gaussian mixture distributions

Table 5 continued

Article	Sentence id	Score	Entities	Sentence
	192	9	Dirichlet distribution (4.0), Comparative:1	For all experiments, we set Dirichlet parameter = 0.01 (prior on word given topic) and Dirichlet parameter = 2/K (prior on topic given document), except where noted

List of scientific papers used in Hypothetical Walkthrough-I

[a]Blei, D. M., & Lafferty, J. D. (2005). *Correlated topic models.* Paper presented at the Proceedings of the 18th International Conference on Neural Information Processing Systems

[b]Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models.* Paper presented at the Proceedings of the 23rd international conference on Machine learning

[c]Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). *Latent dirichlet allocation.* Paper presented at the Advances in neural information processing systems

[d]Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). *Hierarchical topic models and the nested chinese restaurant process.* Paper presented at the Advances in neural information processing systems

[e]Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the national academy of sciences, 101*(suppl 1), 5228-5235

[f]Hofmann, T. (1999). *Probabilistic latent semantic analysis.* Paper presented at the Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence

[g]Minka, T., & Lafferty, J. (2002). *Expectation-propagation for the generative aspect model.* Paper presented at the Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence

[h]Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences, 61*(2), 217-235

[i]Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). *Fast collapsed gibbs sampling for latent dirichlet allocation.* Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining

[j]Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). *Sharing clusters among related groups: Hierarchical Dirichlet processes.* Paper presented at the Advances in neural information processing systems

[k]Teh, Y. W., Newman, D., & Welling, M. (2007). *A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.* Paper presented at the Advances in neural information processing systems

Extractive summary generated in Hypothetical Walkthrough-II

Article	Sentence id	Score	Entities	Sentence
Italian invasion of Ethiopia — 1935	3	16.6	Ethiopia(1), Nazi Germany(9), Article X of the Covenant of the League of Nations(1), Austria(1), Kingdom of Italy(1), Comparative:1	Both Italy and Ethiopia were member nations, but the League did nothing when the former clearly violated the League's Article X.<33> Germany was the only major European nation to openly support the invasion. Italy subsequently dropped its objections to Germany's goal of absorbing Austria.<34>
	2	10.14285	Ethiopia(0.5), Italian Somaliland(1), Kingdom of Italy(0.5), Italian East Africa(1), League of Nations(2), Ethiopian Empire(1), East African Campaign(2), Comparative:1	The war began with the invasion of the Ethiopian Empire (also known as Abyssinia) by the armed forces of the Kingdom of Italy (Regno d'Italia), which was launched from Italian Somaliland and Eritrea.<32> The war resulted in the military occupation of Ethiopia and its annexation into the newly created colony of Italian East Africa (Africa Orientale Italiana, or AOI); in addition it exposed the weakness of the League of Nations as a force to preserve peace
	0	10	Benito Mussolini(5), Comparative:0	Benito Mussolini inspecting troops during the Italo-Ethiopian War, 1935
	1	3	October 1935(1), May 1936(1), Comparative:0	The Second ItaloEthiopian War was a brief colonial war that began in October 1935 and ended in May 1936

Article	Sentence id	Score	Entities	Sentence
Spanish Civil War—1936	4	28	Nazi Germany(9), Union of Soviet Socialist Republics(9), Comparative:1	Both Germany and the USSR used this proxy war as an opportunity to test in combat their most advanced weapons and tactics
	1	16.25	Francisco Franco(1), Benito Mussolini(5), Second Spanish Republic(1), Adolf Hitler(6), Comparative:0	When civil war broke out in Spain, Hitler and Mussolini lent military support to the Nationalist rebels, led by General Francisco Franco
	5	9.125	World War II(4), Francisco Franco(0.5), Francoist Spain(1), April 1939(1), Comparative:1	The Nationalists won the civil war in April 1939; Franco, now dictator, remained officially neutral during World War II
	2	7.5	Second Spanish Republic(0.5), Union of Soviet Socialist Republics(4.5), Comparative:0	The Soviet Union supported the existing government, the Spanish Republic
	0	3	Spanish Civil War(1), Bombing of Guernica(1), Comparative:0	The bombing of Guernica in 1937, during the Spanish Civil War, sparked Europe-wide fears that the next war would be based on bombing of cities with very high civilian casualties
Japanese invasion of China—1937	1	29.57142857	Marco Polo Bridge Incident(1), China(6), Germany(3), Beijing(1), Empire of Japan(4), Union of Soviet Socialist Republics(9), July 1937(1), Comparative:1	In July 1937, Japan captured the former Chinese imperial capital of Peking after instigating the Marco Polo Bridge Incident, which culminated in the Japanese campaign to invade all of China.<37> The Soviets quickly signed a non-aggression pact with China to lend materiel support, effectively ending China's prior co-operation with Germany

Article	Sentence id	Score	Entities	Sentence
	5	21.9	Wuhan(1), Chinese people(5), China(3.0), Xuzhou(1), Chongqing(1), Imperial Japanese Army(2), Empire of Japan(2.0), June 1938(1), Yellow River(1), Kuomintang(2), Comparative:1	In March 1938, Nationalist Chinese forces won their first major victory at Taierzhuang but then the city of Xuzhou was taken by Japanese in May.<44> In June 1938, Chinese forces stalled the Japanese advance by flooding the Yellow River; this manoeuvre bought time for the Chinese to prepare their defences at Wuhan, but the city was taken by October.<45> Japanese military victories did not bring about the collapse of Chinese resistance that Japan had hoped to achieve; instead the Chinese government relocated inland to Chongqing and continued the war.<46><47>
	2	12.25	Communist Party of China(2), Shanghai(1), Generalissimo(2), Taiyuan(1), National Revolutionary Army(1), Empire of Japan(1.0), Pingxing Pass(1), Battle of Shanghai(1), Comparative:1	From September to November, the Japanese attacked Taiyuan,<38><39> as well as engaging the Kuomintang Army around Xinkou<38><39> and Communist forces in Pingxingguan.<40><41> Generalissimo Chiang Kai-shek deployed his best army to defend Shanghai, but, after 3 months of fighting, Shanghai fell.
	4	4	China(1.5), Empire of Japan(0.5), Nanjing(1), Comparative:0	After the fall of Nanking, tens of thousands if not hundreds of thousands of Chinese civilians and disarmed combatants were murdered by the Japanese.<42><43>
	0	3.25	Imperial Japanese Army(1.0), Battle of Shanghai(0.5), Comparative:1	Japanese Imperial Army soldiers during the Battle of Shanghai, 1937

Article	Sentence id	Score	Entities	Sentence
War breaks out in Europe—1939	11	50.53846	France(5), League of Nations(2), Soviet invasion of Poland(1), Nazi Germany(9), November 1939(2), Finland(3), Winter War(1), Red Army(6), United Kingdom(5), Baltic Sea(1), Union of Soviet Socialist Republics(9), Latvia(1), March 1940(1), Comparative:1	After signing the GermanSoviet Treaty of Friendship, Cooperation and Demarcation, the Soviet Union forced the Baltic countriesEstonia, Latvia and Lithuaniato allow it to station Soviet troops in their countries under pacts of "mutual assistance".<76><77><78> Finland rejected territorial demands, prompting a Soviet invasion in November 1939.<79> The resulting Winter War ended in March 1940 with Finnish concessions.<80> Britain and France, treating the Soviet attack on Finland as tantamount to its entering the war on the side of the Germans, responded to the Soviet invasion by supporting the USSR's expulsion from the League of Nations.<78>
	6	31.8	Home Army(1), Polish Underground State(1), France(2.5), Nazi Germany(4.5), Poland(3), United States(7), Romania(5), Enigma machine(1), Polish resistance movement in World War II(1), Baltic states(2), Comparative:1	After the defeat of Poland's armed forces, the Polish resistance established an Underground State and a partisan Home Army.<69> About 100,000 Polish military personnel were evacuated to Romania and the Baltic countries; many of these soldiers later fought against the Germans in other theatres of the war.<70> Poland's Enigma codebreakers were also evacuated to France.<71>
	1	27	Canada(2), New Zealand(2), France(1.25), Poles(1), Australia(7), 1890 British Ultimatum(1), Nazi Germany(2.25), South Africa(1), Commonwealth of Nations(2), Dominion(1), United Kingdom(2.5), Invasion of Poland(1), Comparative:1	On 1 September 1939, Germany invaded Poland under the false pretext that the Poles had carried out a series of sabotage operations against German targets near the border.<63> Two days later, on 3 September, after a British ultimatum to Germany to cease military operations was ignored, Britain and France, followed by the fully independent Dominions<64> of the British Commonwealth<65>Australia (3 September), Canada (10 September), New Zealand (3 September), and South Africa (6 September)declared war on Germany.

Article	Sentence id	Score	Entities	Sentence
	7	20.42857	France(0.625), Adolf Hitler(6), Nazi Germany(1.125), Neville Chamberlain(2), Poland(1.5), United Kingdom(1.25), Union of Soviet Socialist Republics(4.5), Comparative:1	On 6 October, Hitler made a public peace overture to Britain and France, but said that the future of Poland was to be determined exclusively by Germany and the Soviet Union
	4	14.78571	Nazi Germany(0.5625), Poland(0.75), Japan(6), Union of Soviet Socialist Republics(2.25), Polish Armed Forces(1), Invasion of Poland(0.5), Warsaw(1), Comparative:1	On 17 September 1939, after signing a cease-fire with Japan, the Soviets invaded Poland from the east.<68> The Polish army was defeated and Warsaw surrendered to the Germans on 27 September with final pockets of resistance surrendering on 6 October
Western Europe | Mediterranean—1940	0	38.69230	Winston Churchill(3), United Kingdom(5), Norway(1), Sweden(1), Narvik(1), Denmark(1), Allies of World War II(7), Neville Chamberlain(2), May 1940(1), April 1940(1), Norwegian Campaign(2), Prime Minister of the United Kingdom(1), Nazi Germany(9), Comparative:1	In April 1940, Germany invaded Denmark and Norway to protect shipments of iron ore from Sweden, which the Allies were attempting to cut off by unilaterally mining neutral Norwegian waters.<86> Denmark capitulated after a few hours, and despite Allied support, during which the important harbour of Narvik temporarily was recaptured from the Germans, Norway was conquered within 2 months.<87> British discontent over the Norwegian campaign led to the replacement of the British Prime Minister, Neville Chamberlain, with Winston Churchill on 10 May 1940.<88>
	19	29	Ion Antonescu(1), Romania(5), Hungary(3), Axis powers(4), Tripartite Pact(2), Union of Soviet Socialist Republics(9), Comparative:1	The Tripartite Pact stipulated that any country, with the exception of the Soviet Union, not in the war which attacked any Axis Power would be forced to go to war against all three.<109> The Axis expanded in November 1940 when Hungary, Slovakia and Romania joined the Tripartite Pact.<110> Romania would make a major contribution (as did Hungary) to the Axis war against the USSR, partially to recapture territory ceded to the USSR, partially to pursue its leader Ion Antonescu's desire to combat communism.<111>

Article	Sentence id	Score	Entities	Sentence
	4	26.875	Paris(1), Armistice(1), United Kingdom(2.5), Vichy France(2), France(5), Italy(6), Battle of France(1), Nazi Germany(4.5), Comparative:1	On 10 June, Italy invaded France, declaring war on both France and the United Kingdom.<96> Paris fell to the Germans on 14 June and eight days later France signed an armistice with Germany and was soon divided into German and Italian occupation zones,<97> and an unoccupied rump state under the Vichy Regime, which, though officially neutral, was generally aligned with Germany
	12	22	United States(7), China(6), Allies of World War II(3.5), Comparative:0	Throughout this period, the neutral United States took measures to assist China and the Western Allies.
	23	22	United Kingdom(1.25), Benito Mussolini(5), October 1940(1), Battle of Crete(1), Italy(3.0), Adolf Hitler(6), Greece(2), Comparative:0	In October 1940, Italy started the Greco-Italian War because of Mussolini's jealousy of Hitler's success but within days was repulsed with few territorial gains and a stalemate soon occurred.<112> The United Kingdom responded to Greek requests for assistance by sending troops to Crete and providing air support to Greece
Axis attack on the USSR | War breaks out in the Pacific—1941	35	54.33333	United States(7), United Kingdom(5), Theodore Roosevelt(1), China(6), Australia(7), Axis powers(4), Germany(3), Japan(6), Union of Soviet Socialist Republics(9), Comparative:1	These attacks led the United States, United Kingdom, China, Australia and several other states to formally declare war on Japan, whereas the Soviet Union, being heavily involved in large-scale hostilities with European Axis countries, maintained its neutrality agreement with Japan.<166> Germany, followed by the other Axis states, declared war on the United States<167> in solidarity with Japan, citing as justification the American attacks on German war vessels that had been ordered by Roosevelt.<125><168>

Article	Sentence id	Score	Entities	Sentence
	34	40	American Athletic Conference(3), United States(3.5), Pacific Ocean(2), Battle of Hong Kong(1), Southeast Asia(3), Colonialism(1), Attack on Pearl Harbor(1), Malaysia(2), Asia(2), Thailand(2), Philippines(4), Allies of World War II(7), Ontario Highway 7(1), United States Pacific Fleet(1), Empire of Japan(4), Comparative:0	Japan planned to rapidly seize European colonies in Asia to create a large defensive perimeter stretching into the Central Pacific; the Japanese would then be free to exploit the resources of Southeast Asia while exhausting the over-stretched Allies by fighting a defensive war.<163> To prevent American intervention while securing the perimeter it was further planned to neutralise the United States Pacific Fleet and the American military presence in the Philippines from the outset.<164> On 7 December 1941 (8 December in Asian time zones), Japan attacked British and American holdings with near-simultaneous offensives against Southeast Asia and the Central Pacific.<165> These included an attack on the American fleet at Pearl Harbor, the Philippines, landings in Thailand and Malaya<165> and the battle of Hong Kong
	18	31.28571	Eastern Europe(1), Kwantung Army(2), Axis powers(2.0), Red Army(6), Nazi Germany(9), Empire of Japan(2.0), Union of Soviet Socialist Republics(4.5), Comparative:1	By early December, freshly mobilised reserves<143> allowed the Soviets to achieve numerical parity with Axis troops.<144> This, as well as intelligence data which established that a minimal number of Soviet troops in the East would be sufficient to deter any attack by the Japanese Kwantung Army,<145> allowed the Soviets to begin a massive counter-offensive that started on 5 December all along the front and pushed German troops 100250 kilometres (62155 mi) west.<146>
	15	27	Mediterranean Sea(3), Atlantic Charter(2), United States(1.75), United Kingdom(2.5), France(5), Persian Corridor(1), Eastern Front(2), Iran(1), Axis powers(1.0), Germany(1.5), Petroleum industry in Iran(1), Union of Soviet Socialist Republics(2.25), Comparative:1	The diversion of three quarters of the Axis troops and the majority of their air forces from France and the central Mediterranean to the Eastern Front<133> prompted Britain to reconsider its grand strategy.<134> In July, the UK and the Soviet Union formed a military alliance against Germany<135> The British and Soviets invaded neutral Iran to secure the Persian Corridor and Iran's oil fields.<136> In August, the United Kingdom and the United States jointly issued the Atlantic Charter.<137>

Article	Sentence id	Score	Entities	Sentence
	9	22.64062	White movement(1), Caspian Sea(1), Ukraine(3), Finland(3), Baltic states(2), Hungary(3), Adolf Hitler(6), Union of Soviet Socialist Republics(1.125), Comparative:0	They were joined shortly by Finland and Hungary.<125> The primary targets of this surprise offensive<126> were the Baltic region, Moscow and Ukraine, with the ultimate goal of ending the 1941 campaign near the Arkhangelsk-Astrakhan line, from the Caspian to the White Seas
Axis advance stalls: Pacific | Eastern Front | Western Europe/ Atlantic and Mediterranean—1942	26	65.05	American Athletic Conference(3), Atlantic Ocean(1), United States(7), Allies of World War II(7), Battle of Gazala(1), Dieppe Raid(1), Second Battle of El Alamein(1), Libya(2), Australia(7), Japan(6), Europe(4), Operation Torch(1), Kriegsmarine(2), Operation Crusader(1), Axis powers(4), Madagascar(1), Italy(6), Continental Europe(1), Germany(3), Egypt(2), Comparative:1	Exploiting poor American naval command decisions, the German navy ravaged Allied shipping off the American Atlantic coast.<199> By November 1941, Commonwealth forces had launched a counter-offensive, Operation Crusader, in North Africa, and reclaimed all the gains the Germans and Italians had made.<200> In North Africa, the Germans launched an offensive in January, pushing the British back to positions at the Gazala Line by early February,<201> followed by a temporary lull in combat which Germany used to prepare for their upcoming offensives.<202> Concerns the Japanese might use bases in Vichy-held Madagascar caused the British to invade the island in early May 1942.<203> An Axis offensive in Libya forced an Allied retreat deep inside Egypt until Axis forces were stopped at El Alamein.<204> On the Continent, raids of Allied commandos on strategic targets, culminating in the disastrous Dieppe Raid,<205> demonstrated the Western Allies' inability to launch an invasion of continental Europe without much better preparation, equipment, and operational security.<206>

Article	Sentence id	Score	Entities	Sentence
	10	39.42105	Battle of the Java Sea(1), South China Sea(1), 33rd Division(1), Darwin(1), Allies of World War II(3.5), Singapore(1), British Army(1), Indonesia(2), Rabaul(2), Japan(3.0), Battle of Yenangyaung(1), Seamiew Records(1), Indian Ocean(1), Chinese people(5), Thailand(2), Philippines(4), Commonwealth of the Philippines(2), Malaysia(2), Myanmar(2), Comparative:1	By the end of April 1942, Japan and its ally Thailand had almost fully conquered Burma, Malaya, the Dutch East Indies, Singapore, and Rabaul, inflicting severe losses on Allied troops and taking a large number of prisoners.<175> Despite stubborn resistance by Filipino and US forces, the Philippine Commonwealth was eventually captured in May 1942, forcing its government into exile.<176> On 16 April, in Burma, 7,000 British soldiers were encircled by the Japanese 33rd Division during the Battle of Yenangyaung and rescued by the Chinese 38th Division.<177> Japanese forces also achieved naval victories in the South China Sea, Java Sea and Indian Ocean,<178> and bombed the Allied naval base at Darwin, Australia
	13	33.14062	United States(3.5), Alaska(1), Allies of World War II(1.75), Aleutian Islands(2), China(6), Americans(3), Japan(1.5), Chinese people(2.5), Midway Atoll(1), Japanese naval codes(1), Battle of Midway(1), Army group(1), Doolittle Raid(1), Battle of the Coral Sea(1), Imperial Japanese Navy(2), US Marines(1), Comparative:1	The planned invasion was thwarted when an Allied task force, centred on two American fleet carriers, fought Japanese naval forces to a draw in the Battle of the Coral Sea.<181> Japan's next plan, motivated by the earlier Doolittle Raid, was to seize Midway Atoll and lure American carriers into battle to be eliminated; as a diversion, Japan would also send forces to occupy the Aleutian Islands in Alaska.<182> In mid-May, Japan started the Zhejiang-Jiangxi Campaign in China, with the goal of inflicting retribution on the Chinese who aided the surviving American airmen in the Doolittle Raid by destroying air bases and fighting against the Chinese 23rd and 32nd Army Groups.<183><184> In early June, Japan put its operations into action but the Americans, having broken Japanese naval codes in late May, were fully aware of plans and order of battle, and used this knowledge to achieve a decisive victory at Midway over the Imperial Japanese Navy.<185>

Article	Sentence id	Score	Entities	Sentence
	0	27.85937	Atlantic Charter(2), United Kingdom(5), Allies of World War II(0.875), China(3.0), Declaration by United Nations(1), Axis powers(2.0), Four Policemen(1), Union of Soviet Socialist Republics(9), Comparative:1	On 1 January 1942, the Allied Big Four<169>the Soviet Union, China, Britain and the United Statesand 22 smaller or exiled governments issued the Declaration by United Nations, thereby affirming the Atlantic Charter<170>, and agreeing to not to sign a separate peace with the Axis powers
	9	22.875	United States(1.75), Allied invasion of Italy(1), Turkey(2), Allies of World War II(0.4375), France(5), Balkans(3), Americans(1.5), Allied invasion of Sicily(2), Mediterranean Sea(3), Comparative:1	The British and Americans agreed to continue to press the initiative in the Mediterranean by invading Sicily to fully secure the Mediterranean supply routes.<173> Although the British argued for further operations in the Balkans to bring Turkey into the war, in May 1943, the Americans extracted a British commitment to limit Allied operations in the Mediterranean to an invasion of the Italian mainland and to invade France in 1944.<174>
Allies gain momentum—1943	6	65.9	Italy(6), Red Army(6), Adolf Hitler(6), Nazi Germany(9), Hamburg(2), United States(7), Allied invasion of Sicily(2), Union of Soviet Socialist Republics(9), Allies of World War II(7), Benito Mussolini(5), Comparative:1	Within a week, German forces had exhausted themselves against the Soviets' deeply echeloned and well-constructed defences<217> and, for the first time in the war, Hitler cancelled the operation before it had achieved tactical or operational success.<218> This decision was partially affected by the Western Allies' invasion of Sicily launched on 9 July which, combined with previous Italian failures, resulted in the ousting and arrest of Mussolini later that month.<219> Also, in July 1943 the British firebombed Hamburg killing over 40,000 people.<220>

Article	Sentence id	Score	Entities	Sentence
	21	41.92307	Assam(2), Australia(7), Burma Campaign(2), Battle of Kohima(1), Henan(1), United States(3.5), Changsha(4), China(6), Hunan(1), Battle of Imphal(1), Empire of Japan(4), Allies of World War II(3.5), Myitkyina(2), Comparative:1	In March 1944, the Japanese launched the first of two invasions, an operation against British positions in Assam, India,<243> and soon besieged Commonwealth positions at Imphal and Kohima.<244> In May 1944, British forces mounted a counter-offensive that drove Japanese troops back to Burma,<244> and Chinese forces that had invaded northern Burma in late 1943 besieged Japanese troops in Myitkyina.<245> The second Japanese invasion of China aimed to destroy China's main fighting forces, secure railways between Japanese-held territory and capture Allied airfields.<246> By June, the Japanese had conquered the province of Henan and begun a new attack on Changsha in the Hunan province.<247>
	19	33.725	Romania(5), Italy(3.0), Nazi Germany(4.5), Crimean War(2), Axis powers(4), Baltic region(1), Ukraine(3), Union of Soviet Socialist Republics(4.5), Allies of World War II(1.75), Roman Empire(1), Comparative:1	This delay slowed subsequent Soviet operations in the Baltic Sea region.<240> By late May 1944, the Soviets had liberated Crimea, largely expelled Axis forces from Ukraine, and made incursions into Romania, which were repulsed by the Axis troops.<241> The Allied offensives in Italy had succeeded and, at the expense of allowing several German divisions to retreat, on 4 June, Rome was captured.<242>

Article	Sentence id	Score	Entities	Sentence
	15	24.15625	Europe(4), Nazi Germany(2.25), November 1943(1), Burma Campaign(1.0), Joseph Stalin(1), Geography of Japan(1), Winston Churchill(3), Empire of Japan(2.0), NATO(2), Union of Soviet Socialist Republics(2.25), Allies of World War II(0.875), Tehran Conference(1), Comparative:1	By May 1943, as Allied counter-measures became increasingly effective, the resulting sizeable German submarine losses forced a temporary halt of the German Atlantic naval campaign.<230> In November 1943, Franklin D. Roosevelt and Winston Churchill met with Chiang Kai-shek in Cairo and then with Joseph Stalin in Tehran.<231> The former conference determined the post-war return of Japanese territory<232> and the military planning for the Burma Campaign,<233> while the latter included agreement that the Western Allies would invade Europe in 1944 and that the Soviet Union would declare war on Japan within 3 months of Germany's defeat.<234>
	3	21.51562	Canada(2), Gilbert Islands(1), New Zealand(2), Caroline Islands(1), Aleutian Islands(2), Australia(3.5), Pacific Ocean(2), Marshall Islands(1), Empire of Japan(1.0), Allies of World War II(0.4375), Rabaul(2), Chuuk Lagoon(1), Comparative:1	In May 1943, Canadian and US forces were sent to eliminate Japanese forces from the Aleutians.<214> Soon after, the US, with support from Australian and New Zealand forces, began major operations to isolate Rabaul by capturing surrounding islands, and breach the Japanese Central Pacific perimeter at the Gilbert and Marshall Islands.<215> By the end of March 1944, the Allies had completed both of these objectives, and had also neutralised the major Japanese base at Truk in the Caroline Islands

Article	Sentence id	Score	Entities	Sentence
Allies close in—1944	9	48.25	Slovakia(2), Nazi Germany(9), Warsaw Pact(1), Red Army(6), Bulgaria(3), Romania(5), Union of Soviet Socialist Republics(9), Allies of World War II(7), Comparative:1	However, the largest of these in Warsaw, where German soldiers massacred 200,000 civilians, and a national uprising in Slovakia, did not receive Soviet support and were subsequently suppressed by the Germans.<254> The Red Army's strategic offensive in eastern Romania cut off and destroyed the considerable German troops there and triggered a successful coup d'tat in Romania and in Bulgaria, followed by those countries' shift to the Allied side.<255>
	14	30.71428	Chindwin River(1), Assam(2), Myitkyina(2), Chinese people(5), Japan(6), Australia(7), Southeast Asia(3), Comparative:1	By the start of July 1944, Commonwealth forces in Southeast Asia had repelled the Japanese sieges in Assam, pushing the Japanese back to the Chindwin River<260> while the Chinese captured Myitkyina
	2	25	Italy(6), France(5), Nazi Germany(4.5), Liberation of Paris(1), Allies of World War II(3.5), Comparative:1	After reassigning several Allied divisions from Italy, they also attacked southern France.<249> These landings were successful, and led to the defeat of the German Army units in France
	15	24.1	Guilin(1), French Indochina(2), Hengyang(1), China(6), Mount Song(1), Changsha(4), Japan(3.0), Burma Road(1), Guangxi(1), Liuzhou(1), Comparative:1	In September 1944, Chinese force captured the Mount Song to reopen the Burma Road.<261> In China, the Japanese had more successes, having finally captured Changsha in mid-June and the city of Hengyang by early August.<262> Soon after, they invaded the province of Guangxi, winning major engagements against Chinese forces at Guilin and Liuzhou by the end of November<263> and successfully linking up their forces in China and Indochina by mid-December.<264>

Article	Sentence id	Score	Entities	Sentence
	12	19.85714	Douglas MacArthur(1), Siege of Budapest(1), Karelian Isthmus(1), Balkans(3), Finland(3), Union of Soviet Socialist Republics(4.5), Hungary(3), Comparative:1	A few days later, the Soviets launched a massive assault against German-occupied Hungary that lasted until the fall of Budapest in February 1945.<257> Unlike impressive Soviet victories in the Balkans, bitter Finnish resistance to the Soviet offensive in the Karelian Isthmus denied the Soviets occupation of Finland and led to a Soviet-Finnish armistice on relatively mild conditions,<258><259> although Finland was forced to fight their former allies
Axis collapse	Allied victory—1945	6	50	Italy(6), Nazi Germany(9), Berlin(1), Hamburg(2), Nuremberg(1), Poland(3), United States(7), Union of Soviet Socialist Republics(9), Allies of World War II(7), Comparative:0	In early April, the Western Allies finally pushed forward in Italy and swept across western Germany capturing Hamburg and Nuremberg, while Soviet and Polish forces stormed Berlin in late April
	28	34	Sakhalin(1), Manchukuo(1), Kwantung Army(2), Red Army(6), Japan(6), Empire of Japan(4), Union of Soviet Socialist Republics(4.5), Allies of World War II(3.5), Yalta Conference(1), Kuril Islands(1), Comparative:1	The Allies justified the atomic bombings as a military necessity to avoid invading the Japanese home islands which would cost the lives of between 250,000 and 500,000 Allied servicemen and millions of Japanese troops and civilians.<283> Between the two bombings, the Soviets, pursuant to the Yalta agreement, invaded Japanese-held Manchuria, and quickly defeated the Kwantung Army, which was the largest Japanese fighting force.<284><285> The Red Army also captured Sakhalin Island and the Kuril Islands

Article	Sentence id	Score	Entities	Sentence
	25	21.41666	Nazi Germany(4.5), Winston Churchill(3), Prime Minister of Japan(2), United Kingdom(5), Empire of Japan(2.0), Clement Attlee(1), Comparative:1	They confirmed earlier agreements about Germany,<280> and reiterated the demand for unconditional surrender of all Japanese forces by Japan, specifically stating that "the alternative for Japan is prompt and utter destruction".<281> During this conference, the United Kingdom held its general election, and Clement Attlee replaced Churchill as Prime Minister.<282>
	15	20.8	Battle of Leyte(2), Philippines(4), Pacific War(5), United States(3.5), Commonwealth of the Philippines(2), Comparative:1	In the Pacific theatre, American forces accompanied by the forces of the Philippine Commonwealth advanced in the Philippines, clearing Leyte by the end of April 1945
	1	12.66666	Antwerp(1), Nazi Germany(2.25), Italian Campaign(3), Ardennes(1), Allies of World War II(1.75), Western Front(1), Comparative:1	On 16 December 1944, Germany made a last attempt on the Western Front by using most of its remaining reserves to launch a massive counter-offensive in the Ardennes and along the FrenchGerman border to split the Western Allies, encircle large portions of Western Allied troops and capture their primary supply port at Antwerp to prompt a political settlement.<266> By January, the offensive had been repulsed with no strategic objectives fulfilled.<266> In Italy, the Western Allies remained stalemated at the German defensive line

References

Abdul-Rahman, A., Lein, J., Coles, K., Maguire, E., Meyer, M., Wynne, M., et al. (2013). *Rule-based visual mappings–with a case study on poetry visualization.* Paper presented at the Computer Graphics Forum.

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., & Gleicher, M. (2014). *Serendip: Topic model-driven visual exploration of text corpora.* Paper presented at the visual analytics science and technology (VAST), 2014 IEEE conference on.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). *Nymble: a high-performance learning name-finder.* Paper presented at the proceedings of the fifth conference on applied natural language processing.

Blei, D. M., & Lafferty, J. D. (2005). *Correlated topic models.* Paper presented at the Proceedings of the 18th International Conference on Neural Information Processing Systems.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222.

Borthwick, A., & Grishman, R. (1999). *A maximum entropy approach to named entity recognition.* Citeseer.

Bostock, M. (2016). Force-Directed Graph. https://bl.ocks.org/mbostock/4062045. Accessed 8 June 2018.

Bostock, M. (2017). Narrative Charts. Retrieved from https://bl.ocks.org/drzax/81fff35393fb65255621fd0ab8d11bd7. Accessed 8 June 2018.

Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155–205.

Carbonell, J., & Goldstein, J. (1998). *The use of MMR, diversity-based reranking for reordering documents and producing summaries.* Paper presented at the proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval.

Chavalarias, D., & Cointet, J.-P. (2013). Phylomemetic patterns in science evolution—The rise and fall of scientific fields. *PLoS ONE, 8*(2), e54847.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5303–5310.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology, 57*(3), 359–377.

Clement, T., Plaisant, C., & Vuillemot, R. (2009). The Story of One: Humanity scholarship with visualization and text analysis. *Relation, 10*(1.43), 8485.

Correll, M., Witmore, M., & Gleicher, M. (2011). *Exploring collections of tagged text for literary scholarship.* Paper presented at the Computer Graphics Forum.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., et al. (2007). *Discovering interesting usage patterns in text collections: Integrating text mining with visualization.* Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the Association for Information Science and Technology, 63*(12), 2351–2369.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22,* 457–479.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2001). *Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth.* Paper presented at the proceedings of the 17th international conference on data engineering.

Hofmann, T. (1999). *Probabilistic latent semantic analysis.* Paper presented at the Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.

Inselberg, A., & Dimsdale, B. (1987). Parallel coordinates for visualizing multi-dimensional geometry. In *Computer graphics 1987* (pp. 25–44). Springer.

Jindal, N., & Liu, B. (2006a). *Identifying comparative sentences in text documents.* Paper presented at the proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval.

Jindal, N., & Liu, B. (2006b). *Mining comparative sentences and relations.* Paper presented at the AAAI.
Kirschner, P. A., Buckingham-Shum, S. J., & Carr, C. S. (2012). *Visualizing argumentation: Software tools for collaborative and educational sense-making*. London: Springer.
Kobourov, S. G. (2012). Spring embedders and force directed graph drawing algorithms. arXiv preprint arXiv:1201.3011.
Koch, S., John, M., Wörner, M., Müller, A., & Ertl, T. (2014). VarifocalReader—in-depth visual analysis of large text documents. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1723–1732.
Lin, H., & Bilmes, J. (2011). *A class of submodular functions for document summarization.* Paper presented at the proceedings of the 49th annual meeting of the association for computational linguistics: Human Language Technologies-Volume 1.
Liu, S., Wu, Y., Wei, E., Liu, M., & Liu, Y. (2013). Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2436–2445.
McCallum, A., & Li, W. (2003). *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.* Paper presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4.
McCurdy, N., Lein, J., Coles, K., & Meyer, M. (2016). Poemage: Visualizing the sonic topology of a poem. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 439–448.
Mihalcea, R., & Tarau, P. (2004). *Textrank: Bringing order into text.* Paper presented at the Proceedings of the 2004 conference on empirical methods in natural language processing.
Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming, 14*(1), 265–294.
Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington.* Technical Report MSR-TR-2005, 101.
Ping, Q., & Chen, C. (2017). LitStoryTeller: An interactive system for visual exploration of scientific papers leveraging named entities and comparative sentences. In *Proceedings of ISSI 2017–The 16th international conference on scientometrics and informetrics*, Wuhan University, China, 1118-1130.
Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). *Fast collapsed gibbs sampling for latent Dirichlet allocation.* Paper presented at the proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining.
Schneider, N., Hwa, R., Gianfortoni, P., Das, D., Heilman, M., Black, A., et al. (2010). Visualizing topical quotations over time to understand news discourse. Technical Report CMU-LTI-01-103, CMU, 2010.
Tanahashi, Y., & Ma, K.-L. (2012). Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2679–2688.
Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.
Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1190–1197.
Wilhelm, T., Burghardt, M., & Wolff, C. (2013). "To See or Not to See"—An interactive tool for the visualization and analysis of shakespeare plays. In R. Franken Wendelstorf, E. Lindinger, & J. Sieck (Eds.), *Kultur und informatik: Visual worlds & interactive spaces* (pp. 175–185). Glückstadt: Verlag Werner Hülsbusch.
Zhu, X., Goldberg, A., Van Gael, J., & Andrzejewski, D. (2007). *Improving diversity in ranking using absorbing random walks.* Paper presented at the Human Language Technologies 2007: The conference of the north american chapter of the association for computational linguistics; Proceedings of the main conference.