Interactive Visual Analytics on TED Talks

Yaqi Qin* ETH CS Department

Tianyi Liu† ETH CS Department

Tianyang Xu‡ ETH CS Department

Aojie Yu§ ETH CS Department

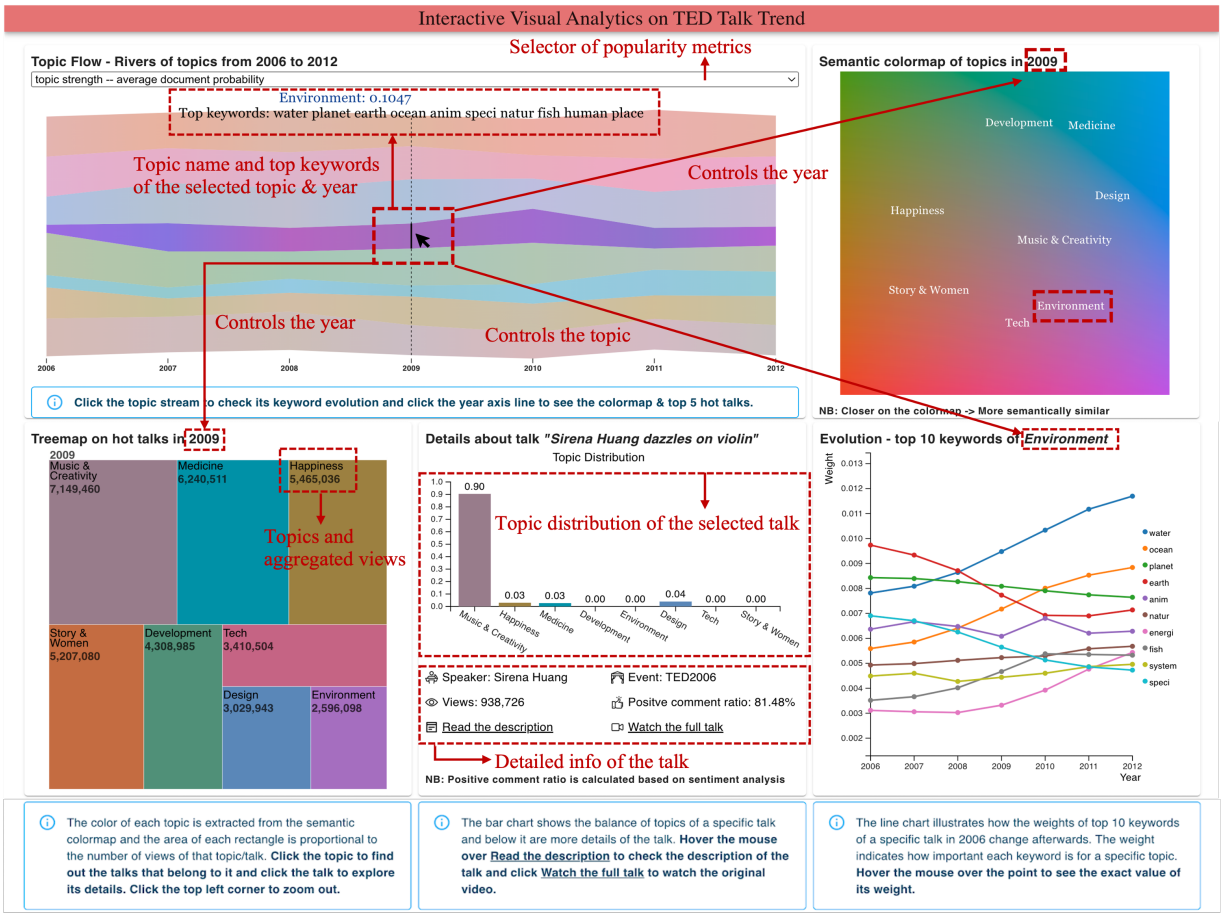

Figure 1: Overview of our interactive dashboard with five components.

Abstract

We present a visual analytic solution that is well-designed to provide investigative functions with fluent interactions to analyze TED talk data. Our work is motivated by the need to keep track of hot topics over time and discover possible factors contributing to the popularity of a talk. Our solution allows the user to view different dimensions of the data at different levels of detail with a mixture of different visualization techniques and smooth interactions. We deploy a Dynamic Latent Dirichlet Allocation model to extract topics from the transcripts of Ted talks and depict the topic evolution over time. Altogether five components are displayed - the topic flow component, the colormap component, the hot talks component, the talk details component, and the keyword evolution component. The user can follow the guidance text below each component to understand the system functionality. Interaction and transition between views are implemented to enable natural actions while performing analysis tasks. The user can interact with the dashboard by hovering, clicking, and using a selector.

Index Terms: Visual analytics—Topic modelling—Text Visualization—Latent Dirichlet Process;

1 Introduction

TED talks are influential videos from expert speakers on education, business, science, technology and creativity, with subtitles in 100+ languages [1]. There have been some attempts on topic modeling and visualization of TED talks to help stakeholders have a better understanding of them, but the analysis of topic popularity and trend is very limited. In this project, we use Dynamic Latent Dirichlet Allocation (LDA) to extract topics from talks' transcripts and Valence Aware Dictionary for sEntiment Reasoning (VADER) to analyze the sentiment of users' comments, which leads to an interactive visual analytic solution to the above-mentinoned issue. The solution allows the user to follow a *topic flow (stream graph)→ colormap*

*e-mail: yaqqin@student.ethz.ch

†e-mail: tianliu@student.ethz.ch

‡e-mail: tianyxu@student.ethz.ch

§e-mail: aojiyu@student.ethz.ch

→ *keyword evolution (line chart)/hot talks (tree map)→ talk details (bar chart)* journey to explore past TED talk trends.

2 Problem Characterization

In this section, we describes the problem solved by our project, including the targeted users, data and tasks.

- **Users and potential uses of the dashboard** In the design stage, we consider two groups of users that might find our project helpful. One group of users are **TED Foundation staff and event holders**, who are familiar with how TED talks are organized. They could make use of the project to keep track of hot topics over time and discover possible factors contributing to the popularity of a talk. The obtained information may prepare them better for future events to attract more audience. The other group of users are **social science researchers** interested in public attention on social issues. Considering the influence of TED talks, the can assist them in analyzing the dynamics of public attention and evaluating the competitiveness of different topics.
- **TED talk data** The TED dataset [5] [6] we use contains 1,149 talks published between 2006 and 2012 from 960 speakers, and 69,023 registered users that have made over 100,000 favorites and 200,000 comments. The talks have data fields including ID, title, description, speaker name, TED event, transcript, publication date, filming date, number of views and user comment threads.
- **Main tasks to complete** Given the users of our interface, the focus of our project is on the analysis of TED talk trend. Specifically, we **[T1] identify hot topics and study how they evolve over time**, **[T2] identify hot talks of each topic**, and **[T3] identify the most important keywords associated with each topic and study the changes in their weights**. We also **[T4] measure the semantic similarity between topics** to facilitate visualization.

3 Data Preprocessing

In the TED dataset, the same talk might be collected multiple times with different IDs, so we first detect and delete duplicate records. Then we extract the raw transcripts of all talks and preprocess them as follows to finally get a vocabulary of 31,347 words.

- **Data cleaning** Punctuation and non-word symbols are removed. We also remove descriptions of audience's reactions which are not related to the contents of talks.
- **POS extraction** In order to compress the vocabulary while keeping the most meaningful words, we conduct Part Of Speech (POS) tag prediction on the tokenized corpus and only preserved nouns, adjectives and verbs.
- **Stop word removal** The most commonly used words in English are removed using the nltk package.
- **Stemming** All words are reduced to its word stem to reduce the redundancy of words with similar meanings.
- **Frequent word removal** Some words like"think","show", and "time" are quite common in TED talks yet meaningless for the inference of the topics, so we further remove the top 40 most frequently occurring words.

For user comment threads, we only keep the first level comments, of which sentiment will be used to rank topics. In other words, all the replies to other comments are excluded, because"the target of their polarity judgment is uncertain" [6].

4 Topic Trend Analysis

In this section, we deploy a Dynamic Latent Dirichlet Allocation model to extract topics from the transcripts of talks, and depicted the topic evolution over time. Based on the results of this dynamic topic model, we come up with three quantitative metrics to evaluate the popularity of each topic and evaluate the similarity between topics by constructing a 2D semantic map. As one of the metrics is dependent on the sentiment of comments, sentiment analysis is also performed.

4.1 Topic Modeling Using Dynamic LDA

Dynamic Latent Dirichlet Allocation (LDA) is a probabilistic model for topic extraction which captures the evolution of topics in a sequentially organized corpus of documents [2]. Here we assume that the talks are generated incrementally at each time step, i.e, talks are grouped by year (of publish date), and each year's talks arise from a set of topics that evolved from last year's topics.

In Dynamic LDA, at each time step t, each topic k is represented as a probability distribution parameterized by β^t over the vocabulary, and each talk d is represented as a probability distribution parameterized by θ^t over all the topics. What makes it dynamic is that β^t and θ^t can vary over time by chaining the distributions with Gaussian noise , as shown below [2].

$$\beta_k^t \mid \beta_k^{t-1} \sim N(\beta_k^{t-1}, \sigma^2 I) \tag{1}$$

$$\alpha^t \mid \alpha^{t-1} \sim N(\alpha^{t-1}, \delta^2 I) \tag{2}$$

$$\theta_d^t \sim N(\alpha^t, a^2 I) \tag{3}$$

where β_k^t represents the vocabulary distribution of topic k at time t, θ_d^t represents the topic distribution of talk k at time t, and α^t models the uncertainty of the proportion of talks that each topic covers at time t.

Therefore, the generative process starts with drawing each topic β_k^t given β_k^{t-1}. Then for each talk, draw θ_d^t given α^t, and sample each word according to the β_k^t of a sampled topic k from the multinomial distribution parameterized by θ_d^t.

Since the number of topics needs to be fixed and preset as a hyperparameter, we run a Hierarchical Dynamic Process (HDP) model beforehand to determine the appropriate number of topics, which is 8.

4.2 Topic Ranking

In order to better interpret the results of the Dynamic LDA model for detailed comparison, we come up with three quantitative metrics to measure topic popularity.

- **Topic Strength (ST)** represents the average topic probability over all talks. A high ST means that on average, a talk is related to the topic with a high probability. The ST score of topic k at time t is defined as:

$$ST_k^t = \frac{1}{|D^t|} \sum_{d \in D^t} P(k \mid d) \tag{4}$$

  where D^t represents the talk corpus at time t and $P(k \mid d)$ represents the probability of talk d belonging to topic k.

- **Topic Coverage (CO)** proposed by Liu et al [4] measures the weighted score of a topic on its average content coverage and its coverage variance. A high CO score means that a topic covers many talks with a high probability, and at the same time, such coverage probability varies a lot among the corpus, which indicates both its prevalence and distinctiveness. The CO score for topic k at time t is formulated as:

$$\mu_k^t = \sum_{d \in D^t} \{P(k \mid d) \times len(d)\} \,/\, \sum_{d \in D^t} len(d) \tag{5}$$

$$\sigma_k^t = \sum_{d \in D^t} \left\{ (P(k \mid d) - \mu_k^t)^2 \times len(d) \right\} / \sum_{d \in D^t} len(d) \quad (6)$$

$$CO_k^t = [\mu_k^t]^{l1} \times [\sigma_k^t]^{l2} \quad (7)$$

where $len(d)$ represents the length of talk d, μ_k^t and σ_k^t represent the mean and standard deviation of the coverage of topic k at time t respectively.

- **Positive Ratio (PR)** represents the average ratio of positive comments under a topic. Here we utilize the sentiment analysis results of comments. A higher PR means that a topic has received a larger percentage of positive comments, indicating that it might be more welcomed by online viewers. The PR score of topic k at time t is formulated as:

$$PR_k^t = \frac{1}{|D^t|} \sum_{d \in D^t} P(k \mid d) \times \frac{|\{C[SENT = positive] \mid d\}}{|\{C \mid d\}|} \quad (8)$$

where $\{C \mid d\}$ represents the set of comments of talk d, and $\{C[SENT = positive] \mid d\}$ represents the set of positive comments.

4.3 Topic Correlation

Since topics are not strictly exclusive from each other, we want to measure their semantic similarity so that at each time step, we can compare the semantic correlation between different topics, and the semantic change of each topic over time. With Dynamic LDA, we get the vocabulary distribution of each topic, which can be regarded as a vector embedding. Afterwards, we perform dimension reduction using t-SNE to reduce the dimensionality of the vector embedding to 2.

As shown in Figure 2, t-SNE successfully preserves the semantic similarity of a topic over seven time steps as a cluster, and is able to distinguish between topics quite well.

4.4 Sentiment Analysis Using VADER

Valence Aware Dictionary for sEntiment Reasoning (VADER) is a rule-based model for general sentiment analysis that makes use of a gold-standard list of lexical features along with their associated sentiment intensity measures [3]. For each comment, the model outputs four scores, namely the positive score, the negative score, the neutral score and the compound score. The last score is computed from the first three and is normalized between -1 and 1, which we use to determine the sentiment of the comment. A higher compound score indicates that the comment is more positive. The general rule is as follows.

$$Comment = \begin{cases} positive, & \text{if compound score} > 0.4 \\ negative, & \text{if compound score} < \text{-}0.4 \\ neutral, & \text{otherwise} \end{cases}$$

5 Visualization

We deploy various visualization techniques in our project using five components, and add both user-component interaction and component interaction.

5.1 Layout Design

We choose dashboard to arrange the five components as shown in Figure 1. Since components interact with each other and the changes in components need to be seen, tabs or scrollytelling are not suitable in our case. The entry point of our visualization is the stream graph on the top left, which also controls the views in the colormap, the treemap and the line chart. The two components on the first row are the basis of the dashboard and the three components below are extensions. Two typical user journeys are *stream graph* $\rightarrow$ *treemap* $\rightarrow$ *talk details* and *stream graph* $\rightarrow$ *keyword evolution*. We add guidance below each component to help users understand the system functionality as well.

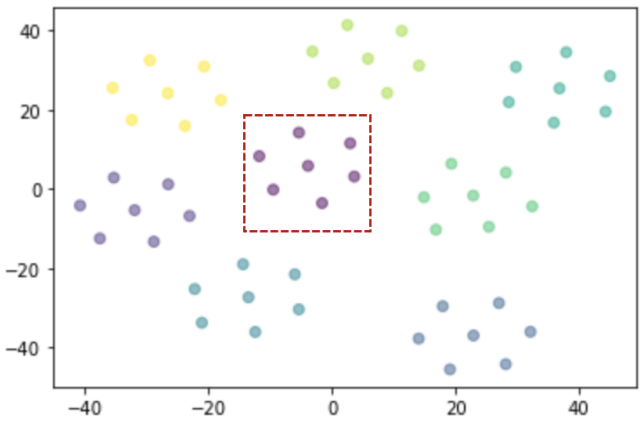

Figure 2: A visualization of the 2D vector representations of eight topics over seven time steps. Each color represents one topic. One cluster corresponding to one topic is highlighted with a red grid.

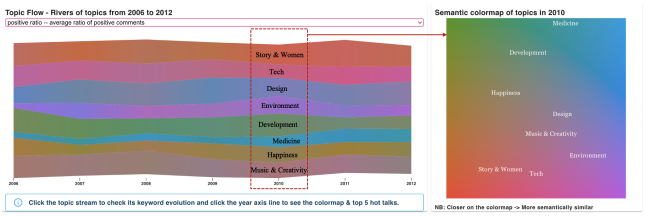

Figure 3: Overview of the topic flow component and the colormap component.

5.2 Topic Flow Component - [T1]

The component we use for visualizing the topic flow over the years is a stream graph, which is the most broadly used graph to visualize the temporal evolution of text data. After performing temporal topic modelling over the TED dataset, we draw a stream graph with 7 time steps (2006-2012) for the user to see the varying popularity represented by the height of each topic stream over the years. As mentioned in section 4.2, three metrics (ST/CO/PR) are used to measure the popularity of topics and by default, we display the graph based on Topic Strength. The selector on the top enables the user to switch between the three metrics.

Interaction with users and other components By using selector, the user could select one quantitative metric out of three (ST/CO/PR) to examine the evolution of topics. By hovering over a specific topic, the user could see that topic stream highlighted, with its popularity measured by the selected metric and the top 10 keywords displayed next to the stream. By clicking on a specific vertical year axis, the user can select a specific year and change the views in other three components (Colormap/Treemap/Line chart).

5.3 Colormap Component - [T4]

The color schemes assigned to each year's hot topics are different, according to the semantic correlation between topics. The color scheme used in the whole dashboard when a specific year is selected is determined by a time-variant colormap. As mentioned in section 4.3, we use t-SNE to reduce the dimensionality of the vector embedding of each topic to 2, then project the 2D embeddings onto a colormap, where we encode semantic similarity as color gradient.

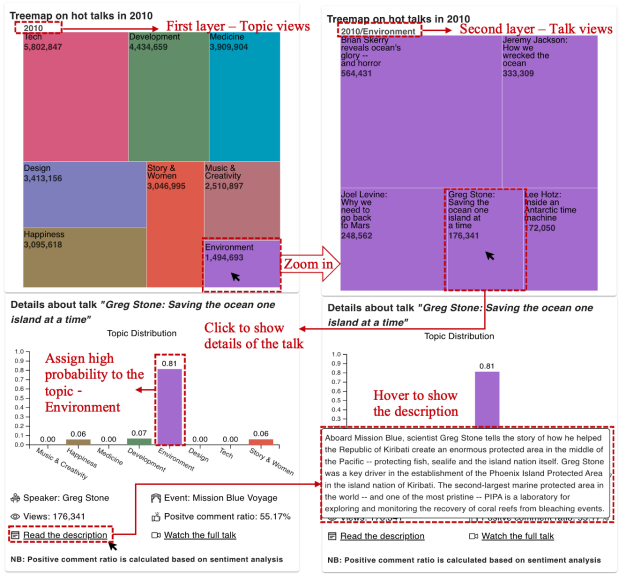

Figure 4: Demonstration of the zoomable treemap

As shown in Figure 3, the topics scatter on the colormap, each corresponding to a color, which is in return used to color the topic stream, the treemap topic block and the bar chart in the talk details component.

Interaction with users and other components After the user clicks on a specific vertical year axis in the stream graph, the colormap will change to the color scheme of the specific year, which is used to color the components mentioned before in the current view. The user can regard the scheme as a more intuitive channel that connects a topic to its visualizations of different aspects.

5.4 Hot Talks Component - [T2]

To associate each topic with the number of views corresponding hot videos have, we use a zoomable treemap in Figure 4 to demonstrate the *year-topic-hot video* hierarchy of the selected year in the stream graph, as treemap is a common hierarchy visualization that emphasizes values of nodes via area encoding. On the first level are the topic blocks, where the area of each block is proportional to the aggregated number of views of each topic. On the second level are the blocks for hot talks, where the area of each block is proportional to the number of views of each talk. All blocks have the talk title or the topic name, the speaker, and the number of views labeled on it.

Interaction with users and other components After the user clicks on a specific vertical year axis in the stream graph, the treemap will change to the view corresponding to that specific year, where the user could compare the popularity of topics directly by checking the area of each topic block. By clicking on a specific topic block, the user will see the second level of the treemap which displays the top 5 hot talks under that topic. By clicking on the upper left prompt, the user could go back to the previous level. By clicking on the block of a specific hot talk, the user could see the details of that talk in the talk details component.

5.5 Talk Details Component - [T2]

In this component we display the details of each talk selected in the treemap, including topic distribution, speaker, TED event it belongs to, number of views and positive comment ratio. The options of reading the description of the talk and watching the full talk are also offered. To visualize the topic distribution, the probabilities that a talk is associated with different topics, a bar chart is used as people tend to correctly estimate lengths.

Interaction with users and other components This component changes as the user clicks on a different talk block in the treemap. By hovering over *Read the description*, the user could see the description of the talk. By clicking on *Watch the full talk*, the user will be directed to the original TED talk page.

5.6 Keyword Evolution Component - [T3]

To visualize how the weights of the top 10 keywords associated with each top in 2006 change over time, we use a line chart with each line corresponding to a keyword, as spatial position ranks high for both expressiveness and effectiveness compared with other visual channels. The standard color scheme consisting of ten categorical colors provided by d3 *schemeCategory10* is chosen to differentiate between keywords.

Interaction with users and other components This component changes as the user clicks on a different topic stream in the stream graph. By hovering over a point in the line chart, the user can see the exact weight of the keyword in a specific year.

6 Case Study

The TED event holders want to compare topic popularity by evaluating the average topic probability over all talks. They first go to the topic flow component and select *topic strength - average document probability* as the quantitative metric using the selector as in Figure 1. Then they hover over a specific topic stream *environment* and over a specific year *2009*. The topic stream of the environment is thus highlighted. The topic strength for the environment - 0.1047, and the top 10 keywords within the environment topic - water, planet, earth, ocean, anim, speci, natur, fish, human, place, are also displayed on top of the stream. Next, they click on the vertical year axis 2009 to change the other four components. The colormap component will display the semantic colormap of topics in 2009, so the TED event holders can observe that the topic *environment* and *tech* are the closest in terms of semantic similarity. Afterwards, they look at the keyword evolution component and find that among the top 10 keywords of the topic environment in 2006, the weight of keyword water constantly increases and has become the largest since 2009. Later, they check the hot talks component to see the number of aggregated views for the topic environment and its coverage compared to those for other topics. By clicking on the environment topic block to zoom in, they see the top 5 hot talks with the most views under the environment topic as in Figure 4. Interested in a certain talk, they click on *Greg Stone: Saving the ocean one island at a time* to change the talk details component and get the basic information about the talk, such as speaker, event, views, positive comment ratio, and the probabilities of this talk belonging to different topics. If they want to know more about the talk, they can further hover over *Read the description* to see the abstract of the talk or even click on *Watch the full talk* to watch this talk.

7 Conclusion and Future Works

In this paper, we have presented an interactive analytical dashboard to help users explore and understand TED talk trend. Our dashboard is useful in four ways. Firstly, it allows users to identify hot topics and study how they evolve over time. Secondly, it enables users to identify hot talks of each topic and explore their details. Thirdly, it helps to identify the most important keywords associated with each topic and study their change. Lastly, it clearly depicts the semantic similarity between topics. Through one case study, we have demonstrated the usability of our techniques in facilitating users to visually analyze the TED talk trend.

Nevertheless, our design also has some limitations. First of all, the current topic analysis only detects and tracts individual topics over time. It does not extract the connection between topics, such as topic splitting and merging. Besides, if the user disagrees with the topic assigned to a certain talk by the dynamic Latent Dirichlet Allocation model, they can not interact with the dashboard to adjust the category and improve the results.

8 Contribution Statements

- **Yaqi Qin**
  - Conducted preprocessing of the transcript data and topic modelling using LDA and HDP
  - Designed and computed the quantitative metrics for topic popularity
  - Modelled topic embeddings and projected them on 2D space using TSNE
  - Designed and implemented the stacked stream graph with interactions, as well as the color map to depict topic similarity
  - Deigned and added interaction among the five components
  - Written the Topic trend analysis part of the report
- **Tianyi Liu**
  - Conducted basic cleaning of the original data and preprocessing of the comment data and sentiment analysis using VADER
  - Participated in the design of the dashboard, added talk details and implemented the design of the dashboard and the line chart
  - Helped with topic modeling using HDP
  - Helped edit the README files every week
  - Helped design the poster, and presented the problem characterization and the machine learning part during the poster session
  - Written the problem characterization part, the sentiment analysis part and the visualization part for talk details and line chart in the report
- **Tianyang Xu**
  - Participated in the design of the dashboard, designed and implemented the zoomable treemap
  - Implemented the interaction between stream graph, colormap, treemap and talk details
  - Helped pushing and tagging weekly commits and editing the README files
  - Written the visualization part for stream graph, colormap and treemap in the report
  - Helped designing the poster, and presented the visualization module during the poster session
- **Aojie Yu**
  - Participated in the design of the dashboard, designed and implemented the bar chart component
  - Written the abstract, introduction, case study, conclusion and future works in the report
  - Designed the poster, and presented it to other groups during the poster session

References

[1] Ted: Ideas worth spreading.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 113–120. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1143844.1143859

[3] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[4] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–28, 2012.

[5] N. Pappas and A. Popescu-Belis. Combining content with user preferences for ted lecture recommendation. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 47–52, 2013. doi: 10.1109/CBMI.2013.6576551

[6] N. Pappas and A. Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 773–776. Association for Computing Machinery, 2013. doi: 10.1145/2484028.2484116