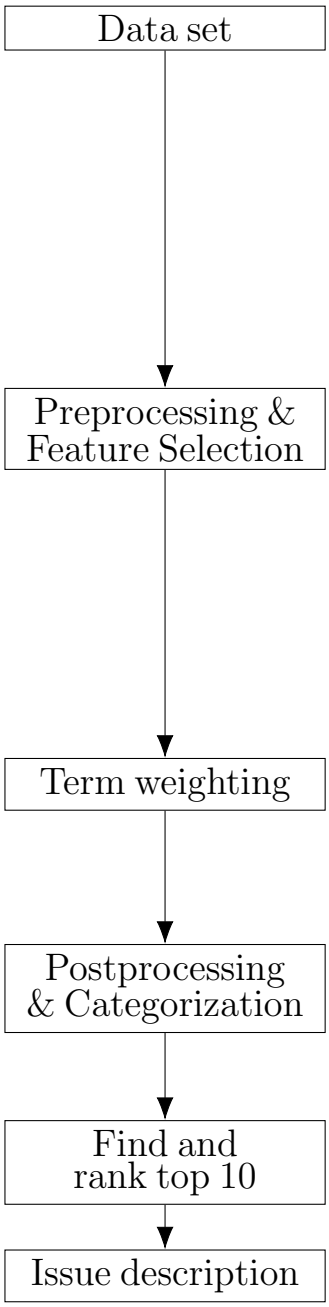
Design Paper

Giang Nguyen, Lydia Kalkbrenner, Philipp Oberwegner

October 29, 2018

I. Issue Trend Analysis



Problem Definition:

Given a collection of newspaper articles for three years, we were asked to find the top 10 issues for each year and rank them based on prominence and salience. Therefore we have to come up with a way to extract issues from the article collection and find a ranking method that can represent the prominence of the issues within the article collection. After ranking the issues we have to extract a description out of the articles that are related to an issue.

Step I: In this step we will create a bag of words for each article and apply tokenization and stemming methods. We will also compare the words to a dictionary and try to match synonyms with WordNet. Then we will need to do some feature selection to reduce the dimension of the vector space model that we are going to create in the next step. As we learned in the lecture, the document frequency measure is a well performing method for feature selection because it is very close to the optimal and also easy to compute, so this could be the the method of our choice.

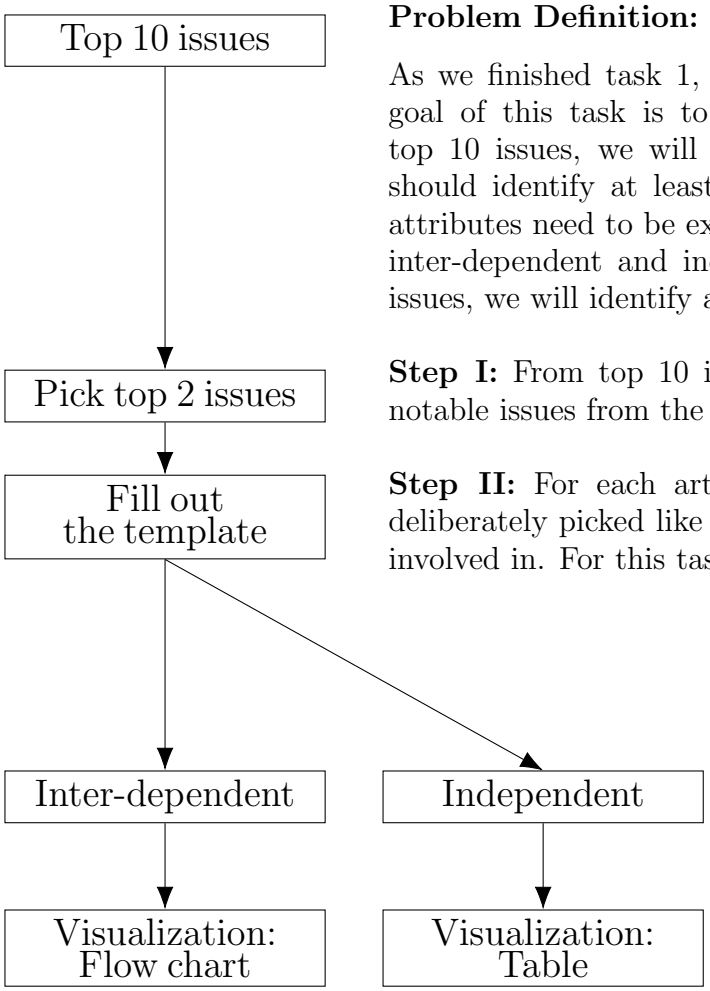
Step II: To transfer the article collection into the vector space model we want to use a more sophisticated measure for term weighting than just frequency based weighting. Because we can't estimate the performance yet, we will try different term weighting methods.

Step III: In this step we will identify the issues by clustering. We make the assumption that the features that we selected occur in the same document if they belong to the same issue. We will identify classes of similar articles by running a clustering algorithm like k-means.

Step IV: To determine the ranking of the identified issue we will use the number of articles that belong to one issue.

Step V: For simplicity we will determine the common words between the documents of an issue for the description of an issue.

II. Issue Tracking



Problem Definition:

As we finished task 1, we now have top 10 issues for each year. The goal of this task is to track events associated with an issue. From top 10 issues, we will pick 2 issues for analyzing, and each issue we should identify at least 10 events. For each event, their outstanding attributes need to be extracted. There are two different types of events: inter-dependent and independent events. For each of the two chosen issues, we will identify all inter-dependent and independent events.

Step I: From top 10 issues for each year, we will pick the two most notable issues from the list of 10.

Step II: For each article, information about article (event) can be deliberately picked like date and place of event, people or organizations involved in. For this task, we may use a domain-specific approach.

Step III: At this step, we should indicate dependence of events of an issue. We assume that inter-dependent incidents will refer to others, and to build the order of events, we could check the chronological relation amongst them. Remaining events, finally, can be group into non-dependent group of issues.

Step IV: As we have two group of non-dependence and dependence, we will visualize the result in table format with non-dependent ones and flow-chart for dependent events.