PoliticalMashup Ngramviewer

Tracking who said what and when in parliament

Bart de Goede
Dispectu
University of Amsterdam
bart@dispectu.com

Justin van Wees
Dispectu
University of Amsterdam
justin@dispectu.com

Maarten Marx
PoliticalMashup
University of Amsterdam
maartenmarx@uva.nl

ABSTRACT

The PoliticalMashup Ngramviewer is an application that allows a user to visualise the use of terms and phrases in the "Tweede Kamer" (the Dutch parliament). Inspired by the Google Books Ngramviewer[1], the PoliticalMashup Ngramviewer additionally allows for *faceting* on politicians and parties, providing a more detailed insight in the use of certain terms and phrases by politicians and parties with different points of view.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION

The Google Books Ngramviewer [2] allows a user to query for phrases consisting of up to 5 terms. The application visualises the relative occurrence of these phrases in a corpus of digitised books written in a specific language over time.

Inspired by the Google Books Ngramviewer, the PoliticalMashup Ngramviewer[2] allows the user to query phrases consisting of up to 7 terms spoken in the Dutch parliament between 1815 and 2012, and visualise the occurrence of those phrases over time. Additionally, the PoliticalMashup Ngramviewer allows the user to facet on politicians and parties, allowing for comparison of the use of phrases through time by parties with different ideologies.

In this demonstration paper we describe the data used in this application, the approach taken with regard to analysing and indexing that data, and examples of how the application could be used in research on agenda setting and linguistics.

2. NGRAMVIEWER

2.1 Data

The PoliticalMashup project [1] aims to make large quantities of political data, such as the proceedings of the Dutch parliament, available and searchable. In addition, a goal of the project is to combine (or *mash up*) political data from different sources, in order to provide for *semantic search*, such as queries for events or persons.

This Ngramviewer is an example of why linking raw text to entities such as persons or parties can be useful: for each word ever uttered in the Dutch parliament, we know who said it, when it was said, to which party that person belonged at that time, and which role that person had at that point in the debate. By linking text to speakers, faceting on persons and parties is enabled.

The data this application uses originates from three sources: Staten-Generaal Digitaal[3], Officiële Bekendmakingen[4] and Parlementair Documentatiecentrum Leiden[5]. PoliticalMashup collected, analysed and transformed data from these sources, determining which speaker said what when, and to which party that speaker belonged at the time. This dataset is freely available via DANS EASY[6].

n-gram	unique terms	without hapaxes
1-grams	2,773,826	992,291
2-grams	38,811,679	12,852,501
3-grams	170,314,738	38,648,440
4-grams	358,360,166	48,621,948
5-grams	498,848,849	36,838,184
6-grams	573,197,917	22,737,318
7-grams	606,867,133	13,655,460
total	2,249,174,308	174,346,142

Table 1: Distribution of unique n-grams in the Ngramviewer corpus for all terms, and with all *hapaxes* (terms that occur only once in the corpus) removed.

[1] http://books.google.com/ngrams
[2] http://ngram.politicalmashup.nl

[3] Project of the Koninklijke Bibliotheek (http://kb.nl/en/), digitising all Dutch parliamentary proceedings between 1814 and 1995 (http://statengeneraaldigitaal.nl/overdezesite).
[4] Portal of the Dutch government, providing a search interface to all govermental proclamations, including parliamentary proceedings since 1995 (https://zoek.officielebekendmakingen.nl/).
[5] Biographical information on politicians and parties (http://www.parlement.com/).
[6] http://www.persistent-identifier.nl/urn:nbn:nl:ui:13-k2g8-5h

DIR 2013, April 26, 2013, Delft, The Netherlands.

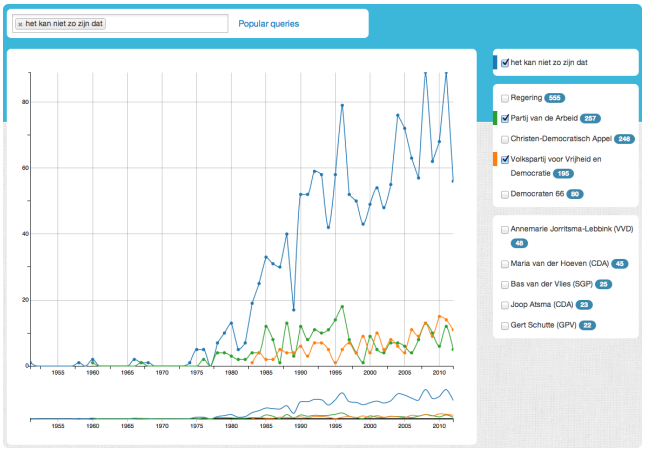

Figure 1: The PoliticalMashup Ngramviewer interface showing results for "het kan niet zo zijn dat", with facets on PvdA and VVD, illustrating the rise of the phrase since the eighties.

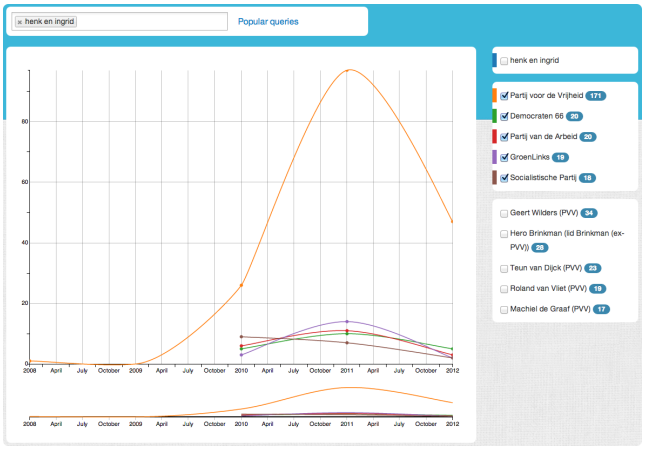

Figure 2: The PoliticalMashup Ngramviewer interface showing results for "Henk en Ingrid", with facets on parties, showing the introduction of the term in 2008, no use in 2009, and that the term is picked up by other parties in 2010.

2.2 Indexing

The PoliticalMashup Ngramviewer is built on top of an Apache Lucene[7] index. We defined a document as *every word* of a specific *politician* spoken on a *particular day*. This allows for comparison of term frequencies per person, per day, which can be aggregated to words spoken by all members of a particular party in a particular time period (week, month, year, etcetera).

We used standard tokenisation and analysis on these documents; lowercasing, character folding and removal of punctuation, but *keeping* stopwords, in order to facilitate search on phrases containing common words such as articles or determiners. Additionally, we constructed word n-grams ($1 \leq n \leq 7$), respecting sentence boundaries.

The index contains data from 4 April 1815 to 9 September 2012, with 326,315 documents (where a document is all the text one person said on one day), 18,572 days for which there are documents, for in total 3,085 politicians which are members of 119 parties or the government. Table 1 shows the distribution of n-grams in the corpus. The second column shows the distribution of n-grams that occur more than once in the corpus, yielding a reduction of the vocabulary size of one order of magnitude. This is partly due to OCR errors (all proceedings predating 1995 are scans of paper archives).

2.3 Architecture

We constructed an inverted index in Lucene, storing the document frequency for each n-gram, and the term frequency for each document that n-gram occurs in.

Additionally, each document has attributes, such as the date the terms of that document were spoken, and identifiers that resolve to politicians and parties[8].

At query time, these identifiers are used to obtain information on persons and parties, which are subsequently cached in a Redis key-value store. This Redis store is also used to cache query results and keep track of popular queries. Also, date frequencies are aggregated to frequencies per year at query time.

[7] `http://lucene.apache.org/core/`

[8] PoliticalMashup maintains a resolver that maps identifiers to persons parties and proceedings.

2.4 Examples

"Het kan niet zo zijn dat"[9] is a popular phrase used by (Dutch) politicians, lending their statement a more urgent feeling, (unconsciously) trying to manipulate their audience, while the person is just ventilating an opinion. Figure 1 shows the rapid increase in use since the eighties, and the use of the Ngramviewer for linguistic research.

"Henk en Ingrid" are a fictional couple, conceived by the Dutch politician Geert Wilders[10], representing the average Dutch family. Figure 2 shows how Wilders' party introduced the phrase in 2008, but was left unused until 2010, when other parties picked up the phrase as well. This example shows the use of the Ngramviewer for agenda-setting.

3. DEMONSTRATION

The demonstration will show how the PoliticalMashup Ngramviewer can be used, displaying a graph of how often the entered phrases occur over time in the proceedings of the Dutch parliament. Also, it will demonstrate faceting on politicians and parties, showing the occurrence of the entered phrases over time for specific politicians and parties.

4. ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 380-52-005 (PoliticalMashup).

5. REFERENCES

[1] M. Marx. Politicalmashup. Retrieved March, 2013 from `http://politicalmashup.nl/over-political-mashup/`.

[2] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[9] In English: "It is unacceptable that ..."

[10] `http://en.wikipedia.org/wiki/Geert_wilders`