Sentiment Analysis of Constitutions

Athina Panotopoulou, Daniel Rockmore, Nick Foti

Computer Science Department, Dartmouth College

The story

We would like to quantitatively measure the **happiness** of written speech.

Our texts are **constitutional preambles** from all over the world.

The method we use is based on the paper [1].

Moreover, we **expand** their method to the **Tf-idf** metric [4].

Input: The LabMT

- **labMT** [3] 1.0: 10222 ranked words.
- Union of 4 sets (10222):
  5000 most frequent words in **Twitter**
  5000 most frequent words in **Google Books**
  5000 most frequent words in **music lyrics**
  5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.

`The ranking is from 1(SAD) to 9(HAPPY) .`

- The ranking is the average of all rankings.

We denote with **h(w)** the estimate of average **happiness** for each word **w** $\in$ $labMT$.

Loading the LabMT

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$. Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731
- $\Delta H = 2$ Number of words: 1008
- $\Delta H = 3$ Number of words: 77
- Using different subsets of labMT highlight different aspects of our data.

Example

If we use words with happiness ranking between 7 and 9, we highlight the positive aspect of a text.

Input: The Dataset

The data set consists of 477 **constitutional preambles** from 171 countries.

Every file name is related to a specific **date**.

477 in total: $84 \in [1787 - 1899]$; $355 \in [1901 - 1999]$; $38 \in [2000 - 2011]$

- Countries that do not exist, or different name.
- Vocabulary that is different from today.
- Translation cannot fully transfer the emotions that has the initial word.

[1] athina.panotopoulou@dartmouth.edu

Preprocessing the Dataset

We are searching for the **exact word**:

"we've", "you've": two distinct words

- Convert all characters to lower case.
- Remove special characters such as : .,?-:
- Replace with gaps.

Computing the happiness

Load the labMT.

Pre-process the texts of our data set C.

`Compute the happiness ranking of each` $c \in C$, $h_{f,avg(c)}$:

- `Create the set of words` $W(c)$ `that are in the preamble` c.
- `Compute the frequency` $f_c(w)$ `for each word` w `in` c.
- `We define` $N(c)$ `as the set of words that are both in` c `and labMT:` $N(c) = W(c) \bigcap labMT$.
- `For each word` w `in` $N(c)$ `we have a rank` $h(w)$.
- `The ranking of the constitutional preamble` c `can then be computed by:` $h_{f,avg(c)} = \frac{\sum_{w \in N(c)} h(w) f_c(w)}{\sum_{z \in N(c)} f_c(z)}$

An extension

We use a different way of ranking the average happiness:

$h_{Tf \times Idf, avg(c)}$

- $|C|$: The size of our data set, the number of constitutional preambles.
- $|C_w|$: The number of constitutional preambles that contain the word w.
- f_c^m : The maximum frequency we have on constitutional preamble c over all words w that belongs to $N(c)$, $f_c^m = max_{w \in N(c)} f_c(w)$.
- $Tf \times Idf_c(w) = \frac{f_c(w)}{f_c^m} \times \log \frac{|C|}{|C_w|}$.

Results: Pearson Correlation Factor

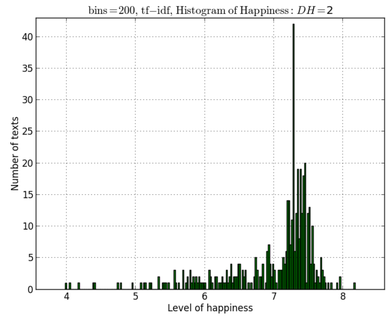
We compute the Pearson Correlation between happiness ranking and other factors [5] —no significant correlation:

Factor	Tf-idf
Limited Government Powers	0.13
Absence of Corruption	0.13
Order and Security	0.18
Fundamental Rights	0.08
Open Government	0.14
Regulatory Enforcement	0.06
Civil Justice	0.1
Criminal Justice	0.09

Results: Histogram

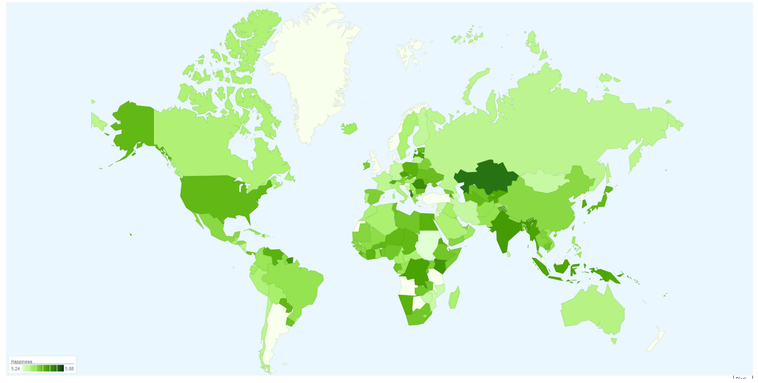
The results tend to lie on the neutral interval, having a positive tendency.

As the ΔH grows the interval grows. It tends to shrink again for $\Delta H = 3$, but now to the left side of the axes-x.



Results: Heatmap

The palest colour represents the lowest happiness score for the specific parameters and the darkest colour the highest.



Based on raw frequency, $\Delta H = 0$.

References

References and Tools:

[1] Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss Christopher M. Danforth, Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter, *http://www.plosone.org/article/info:doi/10.1371/journal.pone.0026752, 2011.*

[2] Google Playground, *code.google.com/apis/ajax/playground/, 2013.*

[3] labMT *http://www.plosone.org.*

[4] Wikipedia, *http://en.wikipedia.org/wiki/Tf%E2%80%93idf, 2013.*

[5] The Rule of Law Index, *http://worldjusticeproject.org/rule-of-law-index, 2013.*

[6] Wikipedia, *http://en.wikipedia.org/wiki/ISO_3166-1, 2013.*