# Homework 2: References

TECHNISCHE UNIVERSITÄT DARMSTADT

- Report (Sample)

## 1 Methodology and Results

### 1.1 Datasets

First, we want to know more about the datasets. Therefore we create some statistics:

| | LGBTQ | Background |
|---|---|---|
| Nr articles | 34,620 | 34,432 |
| Avg (Std) Words/article | 65 (116) | 44 (34) |
| Vocabulary (without stopwords) | 58k | 59k |

Table 1: Dataset statistics

In Figure 3 we can see the time distribution of the articles of both datasets. The number of articles is evenly distributed, with some spikes around 1994, 1999, 2007, 2012 and 2013.

Additionally, we are given the category of each article. This provides some information in which kind of sources the topics related to LGBTQ are discussed. The top-10 categories are `[('U.S.', 8996), ('New York and Region', 6117), ('Opinion', 4432), ('N.Y. / Region', 1875), ('World', 1593), ('Sports', 1007), ('Front Page; U.S.', 897), ('Business', 773), ('Magazine', 707), ('Week in Review', 659)]`.

### 1.2 Context Preprocessing

- read CSV using `pandas.read-csv` with `ISO-8859-1` encoding
- in each text all characters except alphanumerics, whitespace and the '-' symbol were removed (using regex to replace everything except `[a-zA-Z0-9 -]`)
- texts were lowercased, stripped and then splitted at each whitespace

**Analysis** Without removing the seed words from the word cloud, the following graphic is created:


Figure 1: Wordcloud with seed words

Having the seed words removed, we get the fol- lowing word cloud:

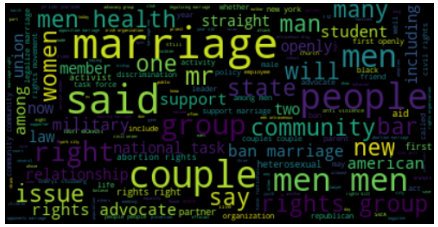



Figure 2: Wordcloud without seed words

From Figure 1 we see that the words from the LGBTQ seed list do often appear together. When we remove the seed words itself (see Figure 2) we get a better picture of the context of LGBTQ articles. One topic which constantly appears in context with LGBTQ-related topics is `marriage` or a form or partnership of people (`couple`, `partner`, `relationship`). Other words are indicating that the articles discuss a group people (e.g. `men`, `women`, `community`, `union`) and might talk about legal points (`right`, `ban`, `law`, `issue`).

- Report (Sample)

| | LGBTQ | Background |
|---|---|---|
| pos. sentiment | 111,315 | 73,320 |
| (unique) | (1,544) | (1,556) |
| neg. sentiment | 210,411 | 138,844 |
| (unique) | (3455) | (3310) |
| pos. percentage | 5% | 5% |
| neg. percentage | 10% | 9% |

Table 2: Sentiment word coverage

### 1.3 Sentiment

Table 2 shows the number of sentiment words contained in both datasets. As we have seen in Table 1, the LGBTQ articles are longer on average. Thus we see both more positive and negative sentiment words in the LGBTQ dataset. With regards to the number of unique sentiment words, the numbers are comparable for both datasets. While the percentage of positive sentiment words is equal for both datasets (5%), the LGBTQ articles contain more negative sentiment words ($10\% > 9\%$).

In Table 3 the top-10 positive and negative sentiment words used in both datasets are noted. Both datasets share a large amount of sentiment words. The positive sentiment words can be considered relevant for people's rights movements (e.g. `right`, `lead`,`support`,`unity`) whereas the negative sentiment words are disrespectful words describing characteristics of people (e.g. `mar`), descriptions of "defects" like `din` and `dent` or words related to bad deeds (e.g. `sin`, `lie`).

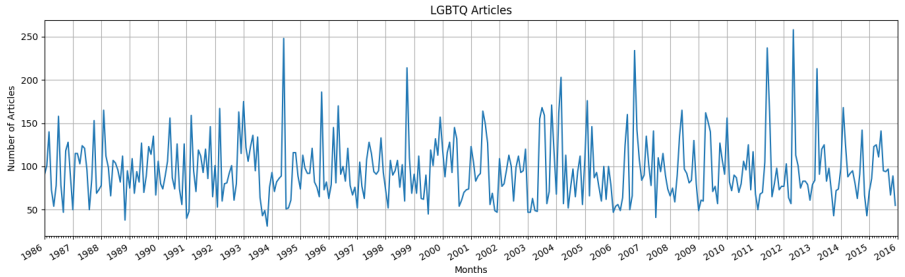



Figure 3: Time Distribution of articles

| | LGBTQ | Background |
|---|---|---|
| top-10 positive sentiment words | right,led,like,gain,support,gains,win,work,lead,unity | led,like,win,work,gain,lead,ease,hot,right,top |
| top-10 negative sentiment words | mar,din,ding,sin,dent,ugh,cons,ire,sue,lie | din,ding,sin,ugh,dent,ire,pan,mar,lie,ail |

Table 3: Sentiment words

- explain **what** you did and **why** it makes sense
- report those results which you actually reference in the text
- interpret your results and draw conclusions!
- report necessary steps and assumptions to reproduce your results

- Confusion Matrix

| | | Actual Label | | |
|---|---|---|---|---|
| | | **Spam** | **No Spam** | |
| **Predicted Label** | **Spam** | 5<br>True Positive | 2<br>False Positive | 7 |
| | **No Spam** | 3<br>False Negative | 10<br>True Negative | 13 |
| | | 8 | 12 | 20 |

Metrics:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ , F1 score = $2 * \frac{P*R}{P+R}$ , False Positive Rate = $\frac{FP}{TN+FP}$
- Precision (P) = $\frac{TP}{TP+FP}$ , Recall (R) = $\frac{TP}{TP+FN}$

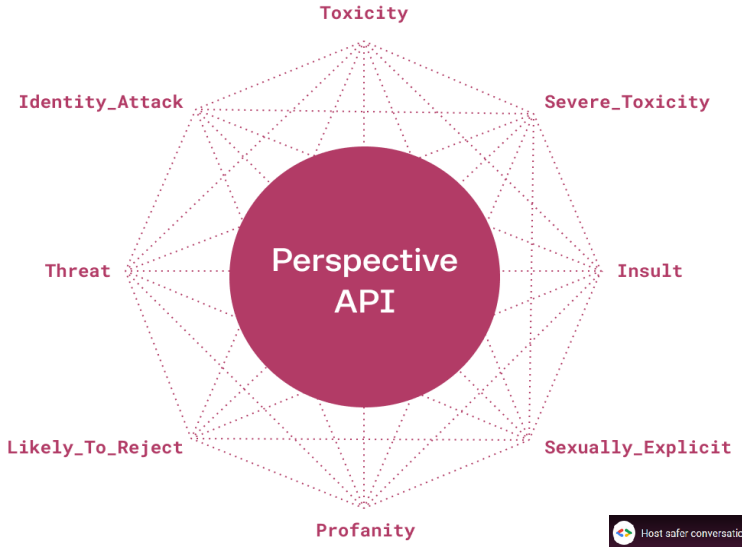
# Perspective API (Google's classifier)

INPUT: TEXT

"Shut up. You're an idiot!"

OUTPUT: SCORE

| | |
|---|---|
| Toxicity | 0.99 |
| Severe_Toxicity | 0.75 |
| Insult | 1.0 |
| Sexually_Explicit | 0.04 |
| Profanity | 0.93 |
| Likely_To_Reject | 0.99 |
| Threat | 0.15 |
| Identity_Attack | 0.03 |





- Website = https://perspectiveapi.com/