The News Classifier

Final Project

Ulises Aquino
and Javier Alvarez Jiménez

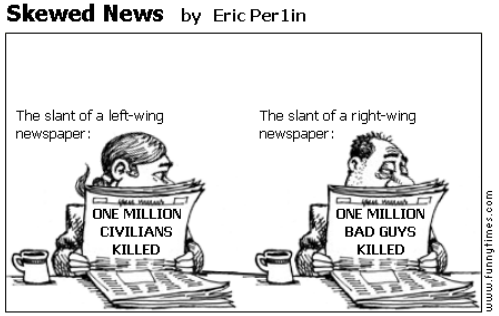
IronHack
Mexico

December 2019

The News

- Nowadays, we have a lot of information, news included.

- The information can be biased.



Steps

- We scrap as many news as we can
- We make a sentiment analysis for all the news (SentimentIntensityAnalyzer from nltk)
- We clusterize the news (vectorization with TfidfVectorizer, clustering with HDBSCAN and cosine similarity)
- We summarize every cluster, and get a sentiment score per cluster
- We upload all the work to mongodb (harder than it sound, ids, structure, etc)

Text Summarization[1]

Algorithm:

- Text $\rightarrow$ sentences
- Get bag of words (avoid stop words, cleaning)
- Give scores to words (in this case, according to the relative frequency)
- From words scores, we can rate the sentences
- We keep the sentences with the highest scores

[1] https://stackabuse.com/text-summarization-with-nltk-in-python/

Recommendation System

We want to build a recommendation System for the article. We use look for the most similar user and recommend articles from it. The key point is to get scores from the users.

- $+1$ if the user read the article
- $+1$ if the user saves the article
- If the user comments the article, we make a sentiment analysis and we add the compound value

Improvements

The clustering system

```
{'title': 'New Zealand police say two bodies may never be found after volcano eruption.',
 'sumary': "New Zealand military team recovers 6 more bodies after White Island volcano eruption, pushing death to
ll to 14 The bodies of two people missing after a volcano erupted on New Zealand's White Island last week likely w
ashed out to sea and may never be found, police said Wednesday. New Zealand police say two bodies may never be fou
nd after volcano eruption. Winona Langford and Hayden Marshall-Inman, missing since the White Island Volcano erupt
ed, were apparently washed out to sea and may never be found. 2 bodies still missing from volcano eruption off New
Zealand that killed 16 people. 'Wait for Mother Nature': Last 2 victims of New Zealand's volcano eruption may neve
r be found. The bodies of six victims of a volcanic eruption on a popular New Zealand tourist island were recovere
d Friday. The volcano's continued venting previously delayed plans by authorities to recover the bodies.",
 'sent':                                  compound
 News_fount
 British Broadcasting Corporation   -0.3182
 Cable News Network                 -0.2960
 USA Today                          -0.2334}
```

Figure: Good cluster.

```
{'title': "Ryan's World is ranked number one for the second year in a row.",
 'sumary': "As President Donald Trump deals with impeachment, here's a look at the people, teams and organizations
in sports who should face impeachment. A case linked to a $670 million acquisition shows that there may be a limit
to the authorities' abuse of law enforcement to advance corrupt business interests. Sarah, 30, says she could have
paid for a holiday with the money she blew on one expensive night out. A year after Cuba's government allowed 3G o
n the island, we follow three activists using the internet to drive rapid change. Ryan's World is ranked number on
e for the second year in a row. A website will feature some of the beloved comic strip's classics and, Larson say
```

Figure: Bad cluster.

- Ideas after use it
- Suggestions...