Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos
Máster Universitario en Ingeniería Informática
Intelligent Systems

Carlos Miguel Alonso
z170243
GiHub Repository
20/06/2022

NLP Text Classification - SpaceNews

Problems

The articles in the Space News data set are not classified by the topic they talk about (transport, military, news, etc.), so by applying an NLP text classifier, we want to automatically classify these articles based on their contents. Furthermore, they also do not provide the keywords of the articles, so we will also solve this using TF-IDF techniques.

This data set contains more than 17.000 articles related to the space industry covering news, commercial, civil, launches, military, and also opinion articles.

Experiments

First, we have discarded the columns we do not want (`url`, `author`, and `date`), since we want to keep the content of the article (`title, content,` and `postexcerpt`) making it a single line (concatenation of these three columns).

Once we have the text of each article, we we check the encoding and normalization of each article and clean it by deleting symbols, extra spaces, and tabs, and lowercasing the full text.

With the data clean, we focus on getting the keywords of each article, but first, we will examine the most common words across all articles. To do this, we create a `corpus` and use it to create a document-feature matrix using the `tm` and `quanteda` R libraries, to analyze these words and find the `topfeatures` of the articles.

We repeat this process but with each article individually, using each `corpus` to create the document-feature matrix and calculate de IDF of each token, since the TF does not give the appropriate words since we take into account the word frequency in other documents aswell.

We have been unable to implement the text classifier we intended since none of the articles are classified by the topic they talk about, but we have setup the code for a future Naive Bayes text classifier if, at some point in the future, the data set includes this classification which allows the training and testing of this model.

Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos
Máster Universitario en Ingeniería Informática
Intelligent Systems

Carlos Miguel Alonso
z170243
GiHub Repository
20/06/2022

Results

In Listings 1 and 2 we see the most and least frequent words of all articles. Notice the most frequent words of all articles are expected, since the articles talk about space related topics, we expected to see such words. On the other hand, the least frequent words are, mainly, names of people and places, which explains why they do not appear more frequently.

```
launch      satellite       nasa        satellites      company
66283       60990           52525       41581           32042
force       commercial      mission     year            million
29253       28916           28587       27753           25162
```

Listing 1: Most frequent words of all articles

```
fatima      cliques     stoking     jaakko      melone      thar
1           1           1           1           1           1
seafood     ipgp        cit         gamaliel
1           1           1           1
```

Listing 2: Least frequent words of all articles

After processing each article, we get their keywords, which give a very brief description of the contents of the articles as shown in Listing 3. These keywords could be added to the articles to give extra information.

```
[1] "astrocast buying hiber boost funding expansion plans tampa fla cash”
[2] "sierra partner spirit aerosystems dream chaser cargo modules
washington announced"
[3] "rocket launch china next station module arrives center helsinki long"
[4] "quad china launches satellite based earth observation initiative brics
nations"
[5] "house committee questions proposed delay nasa asteroid mission
washington members"
```

Listing 3: Keywords of first five articles

Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos
Máster Universitario en Ingeniería Informática
Intelligent Systems

Carlos Miguel Alonso
z170243
GiHub Repository
20/06/2022

Conclusión

Through the use ofTf-IDF, we have been able to get a list of the keywords of each article that could be included on it, giving the reader a brief description of the contents of the articles. These keywords could be used to classify the articles too instead of the whole text, but to do this, more work has to be done on this text classifier.

If we were able to get the articles classified, or at least a small portion of them, big enough to train and test the model, we would be able to prove if a Naive Bayes model is capable of classifying the articles based on their content, but since we lack this article classification, this is not currently possible.

Through this brief project, we were able to understand the current capabilities of NLP frameworks and the importance of properly formatted and classified data for these projects, which is the core of NLP projects, since without data, this field would not advance, at least, not this quickly.