Reuters news classification

December 17, 2017

Data description

The dataset is composed of 7769 training sample and around 3000 test samples. Each new can belong to one or several of the 90 topics, the class are strongly unbalance and some topics only appear a few times. The goal is to predict if a news belong to a topic. The classification algorithm is simple, first a preprocessing is done on the corpus, then a RNN is fitted to classify the different news.

Data preprocessing

The data preprocessing is done with the following process:

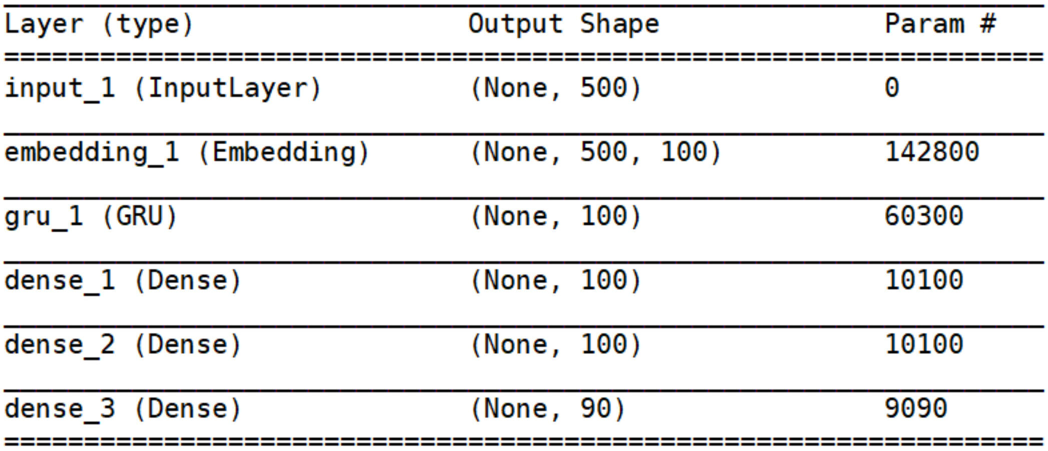
1. Each text is tokenized, it's transformed into a list of word and symbols. The split is done on the spaces.
2. Stops words and punctuation are removed from the tokenized texts. Stops words are mainly conjunctions and logical words, while they can be useful for semantic analysis they are more likely to add noise for topic modeling.
3. Date and numbers are replaced by a placeholder 'datetime'. The topics are not related to an event or to numbers, hence, replacing dates and number by a common placeholder reduce the noise that a large number of different figures and dates would induce.
4. The tokens are stemmed, each token is reduce to its semantical root. This make sense for topic classification.The meaning of the word hold enough information to model the topic and keeping only the root will make the vocabulary smaller (training will be faster and easier to generalize).

Now, each text is represented by a list of tokens. Common words are represented by their stemmed counterpart.

Building the vocabulary and vectorizing the texts

Each word need to be mapped with an integer. A dictionary is built with words that are frequent enough (they are in more than 0.5% of the news). We end up with a dictionary of x words. Using this mapping, each news is transformed into a sequence of integers of length 400. if the initial sequence was shorter, 0 are ended to complete the end of the sequence. If it's longer, the sequence is cut.
Now, we have a numpy array of dimension number of text by sequence length (7769 by 400). It rows contains the sequence of integers representing the stemmed words.

Neural network architecture



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 500)               0
_________________________________________________________________
embedding_1 (Embedding)      (None, 500, 100)          142800
_________________________________________________________________
gru_1 (GRU)                  (None, 100)               60300
_________________________________________________________________
dense_1 (Dense)              (None, 100)               10100
_________________________________________________________________
dense_2 (Dense)              (None, 100)               10100
_________________________________________________________________
dense_3 (Dense)              (None, 90)                9090
=================================================================
```

GRU was used instead of LSTM, performance were similar and switching GRU to LSTM speeded the training.GRU was used with the default parameters (tanh activation function, no regularization nor dropout). Relu activation was used for the dense layer 1 and 2. The last layer uses a sigmoid activation (required for multi-label classification).

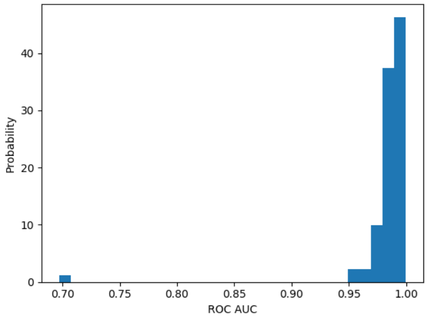
Training and performance

Training

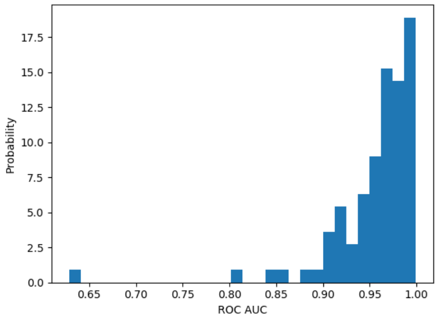
The training was done on the 7769 samples of the training set with a RMSprop optimizer. The NN is trained to minimize the binary cross entropy. The loss stops improving after 12-14 epochs on the the validation sets, the final network is trained for 14 epochs.

Performance

The model is evaluated with using ROC AUC



Repartition of the ROC AUC score for the different topics on the train set (Mean AUC: 0.982)

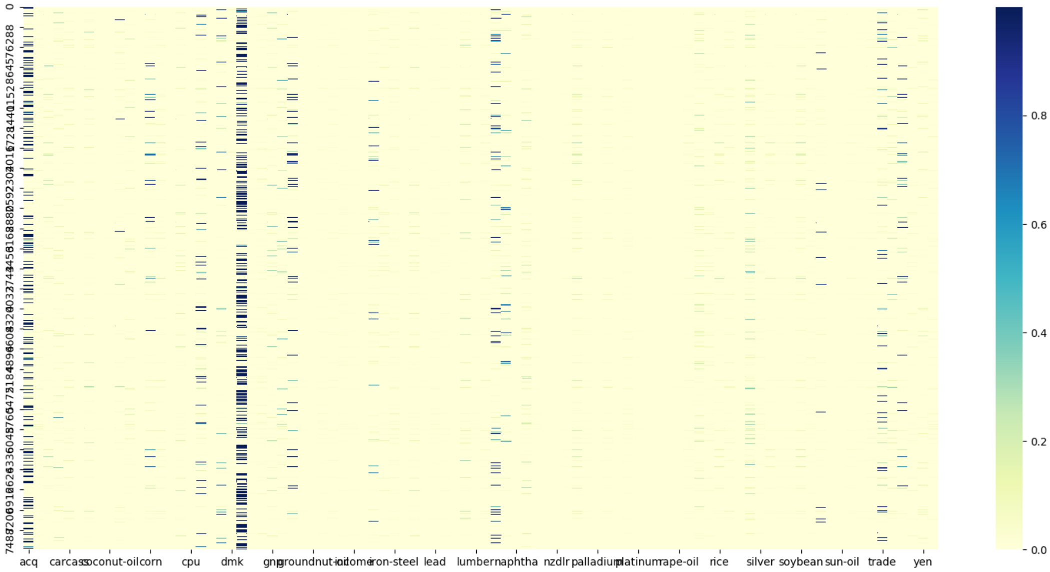


Repartition of the ROC AUC score for the different topics on the train set (Mean AUC: 0.956)

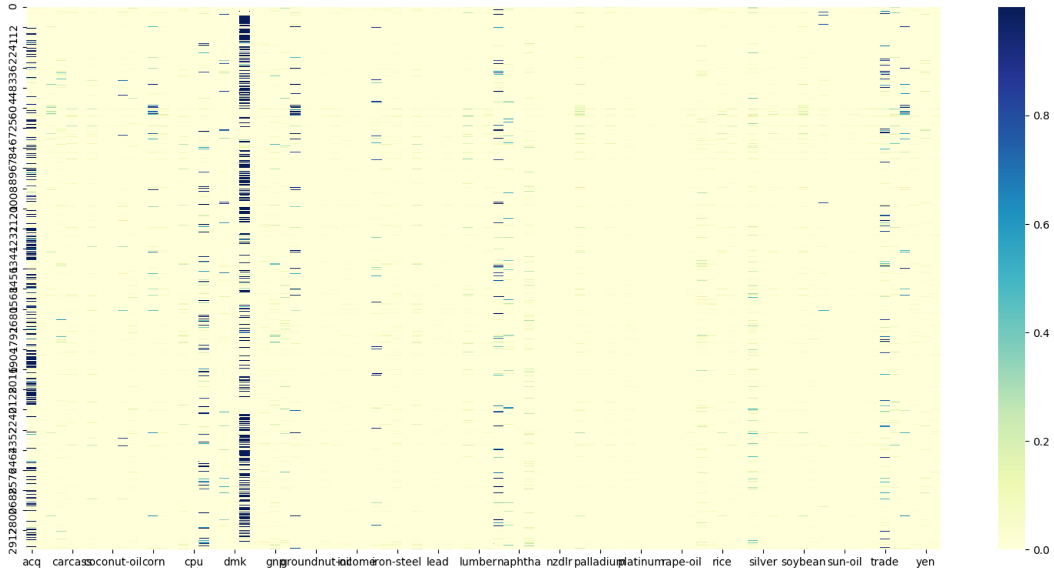
The model shows great ROC AUC and is a clear improvement from random.

Visualization of the laster layer

To visualize the last layer, a heatmap showing the results for the different news was plotted.



Predicted classes on the training set



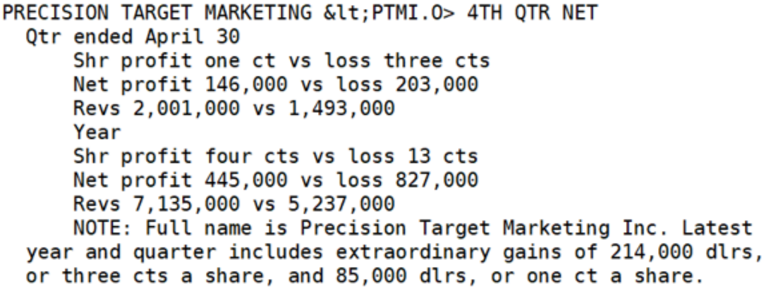
Predicted classes on the validation set

Computing news embedding and proximity

Using the Keras functional API, we can make a forward pass to get the GRU output for each document. The computed vector is a representation of the document in an embedding space. To compute the similarity between two documents, the cosine similarity of their embedded representation was used.

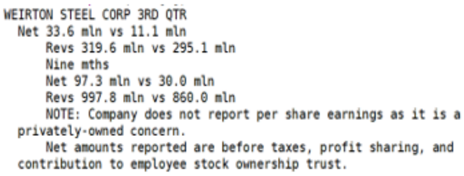
k-th closest news to a given new

Using the cosine similarity defined previously, the 3 closest articles to a given articles can be computed. Here are two examples:



PRECISION TARGET MARKETING <PTMI.O> 4TH QTR NET
Qtr ended April 30
Shr profit one ct vs loss three cts
Net profit 146,000 vs loss 203,000
Revs 2,001,000 vs 1,493,000
Year
Shr profit four cts vs loss 13 cts
Net profit 445,000 vs loss 827,000
Revs 7,135,000 vs 5,237,000

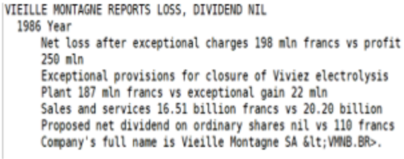
NOTE: Full name is Precision Target Marketing Inc. Latest year and quarter includes extraordinary gains of 214,000 dlrs, or three cts a share, and 85,000 dlrs, or one ct a share.



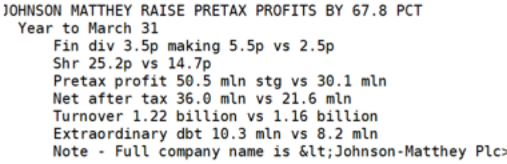
WEIRTON STEEL CORP 3RD QTR
Net 33.6 mln vs 11.1 mln
Revs 319.6 mln vs 295.1 mln
Nine mths
Net 97.3 mln vs 30.0 mln
Revs 997.8 mln vs 860.0 mln

NOTE: Company does not report per share earnings as it is a privately-owned concern.

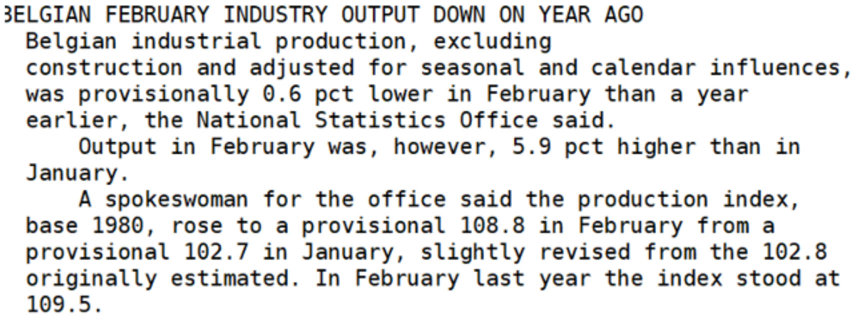
Net amounts reported are before taxes, profit sharing, and contribution to employee stock ownership trust.



VIEILLE MONTAGNE REPORTS LOSS, DIVIDEND NIL
1986 Year
Net loss after exceptional charges 198 mln francs vs profit 250 mln
Exceptional provisions for closure of Viviez electrolysis Plant 187 mln francs vs exceptional gain 22 mln
Sales and services 16.51 billion francs vs 20.20 billion
Proposed net dividend on ordinary shares nil vs 110 francs
Company's full name is Vieille Montagne SA <VMNB.BR>.



JOHNSON MATTHEY RAISE PRETAX PROFITS BY 67.8 PCT
Year to March 31
Fin div 3.5p making 5.5p vs 2.5p
Shr 25.2p vs 14.7p
Pretax profit 50.5 mln stg vs 30.1 mln
Net after tax 36.0 mln vs 21.6 mln
Turnover 1.22 billion vs 1.16 billion
Extraordinary dbt 10.3 mln vs 8.2 mln
Note - Full company name is <Johnson-Matthey Plc>

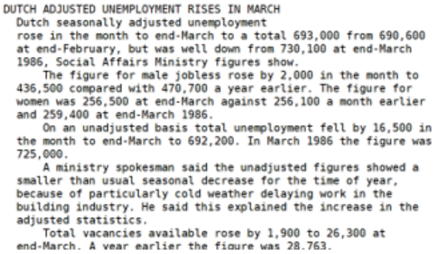


BELGIAN FEBRUARY INDUSTRY OUTPUT DOWN ON YEAR AGO

Belgian industrial production, excluding construction and adjusted for seasonal and calendar influences, was provisionally 0.6 pct lower in February than a year earlier, the National Statistics Office said.

Output in February was, however, 5.9 pct higher than in January.

A spokeswoman for the office said the production index, base 1980, rose to a provisional 108.8 in February from a provisional 102.7 in January, slightly revised from the 102.8 originally estimated. In February last year the index stood at 109.5.



DUTCH ADJUSTED UNEMPLOYMENT RISES IN MARCH

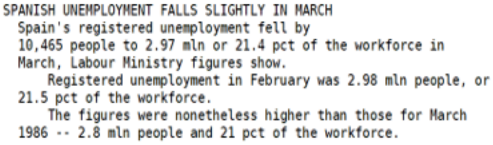
Dutch seasonally adjusted unemployment rose in the month to end-March to a total 693,000 from 690,600 at end-February, but was well down from 730,100 at end-March 1986, Social Affairs Ministry figures show.

The figure for male jobless rose by 2,000 in the month to 436,500 compared with 470,700 a year earlier. The figure for women was 256,500 at end-March against 256,100 a month earlier and 259,400 at end-March 1986.

On an unadjusted basis total unemployment fell by 16,500 in the month to end-March to 692,200. In March 1986 the figure was 725,000.

A ministry spokesman said the unadjusted figures showed a smaller than usual seasonal decrease for the time of year, because of particularly cold weather delaying work in the building industry. He said this explained the increase in the adjusted statistics.

Total vacancies available rose by 1,900 to 26,300 at end-March. A year earlier the figure was 28,763.

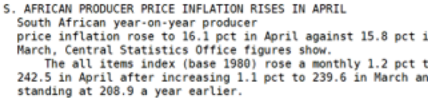


SPANISH UNEMPLOYMENT FALLS SLIGHTLY IN MARCH

Spain's registered unemployment fell by 10,465 people to 2.97 mln or 21.4 pct of the workforce in March, Labour Ministry figures show.

Registered unemployment in February was 2.98 mln people, or 21.5 pct of the workforce.

The figures were nonetheless higher than those for March 1986 -- 2.8 mln people and 21 pct of the workforce.



S. AFRICAN PRODUCER PRICE INFLATION RISES IN APRIL

South African year-on-year producer price inflation rose to 16.1 pct in April against 15.8 pct i
March, Central Statistics Office figures show.

The all items index (base 1980) rose a monthly 1.2 pct t
242.5 in April after increasing 1.1 pct to 239.6 in March an
standing at 208.9 a year earlier.