



Fake-News Detection System

Σταμάτιος Ορφανός
Ε17113



Περιεχόμενα

1. Δεδομένα
2. Προεπεξεργασία Δεδομένων
3. Word2Vec
4. Embedding
5. Αρχιτεκτονική Μοντέλου Κατηγοριοποίησης
6. Αποτελέσματα Εκπαίδευσης
7. Demo



Δεδομένα

Τα δεδομένα για την ανάπτυξη της εφαρμογής και την εκπαίδευση του μοντέλου είναι προέρχονται από την σελίδα :

<https://www.kaggle.com/c/fake-news/data>

Προτού ξεκινήσουμε με την ανάλυση του μοντέλου πρέπει να αναφερθούμε στη μορφή των δεδομένων και την προεπεξεργασία που έγινε έτσι ώστε να μπορούμε να τα χρησιμοποιήσουμε στο μοντέλο.

Δεδομένα

Τα δεδομένα έχουν την εξής μορφή αρχικά :

```
id,title,author,text,label
0,House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It,Darrell Lucas,"House Dem /
With apologies to Keith Olbermann, there is no doubt who the Worst Person in The World is this week-FBI Direc
As we now know, Comey notified the Republican chairmen and Democratic ranking members of the House Intelligen
- Jason Chaffetz (@jasoninthehouse) October 28, 2016
Of course, we now know that this was not the case . Comey was actually saying that it was reviewing the email
But according to a senior House Democratic aide, misreading that letter may have been the least of Chaffetz'
So let's see if we've got this right. The FBI director tells Chaffetz and other GOP committee chairmen about
There has already been talk on Daily Kos that Comey himself provided advance notice of this letter to Chaffet
What it does suggest, however, is that Chaffetz is acting in a way that makes Dan Burton and Darrell Issa loo
Granted, it's not likely that Chaffetz will have to answer for this. He sits in a ridiculously Republican dis
Darrell is a 30-something graduate of the University of North Carolina who considers himself a journalist of
```



Προεπεξεργασία

Όπως μπορούμε να παρατηρήσουμε τα δεδομένα δεν βρίσκονται στην κατάλληλη μορφή επεξεργασίας. Για αυτό το λόγο ακολουθούμε τα παρακάτω βήματα προεπεξεργασίας :

1. Μετατροπή όλων των λέξεων σε πεζά όπου χρειάζεται
2. Αφαίρεση όλων των tags είτε html tags είτε xml tags
3. Αφαίρεση όλων των σημείων στίξης
4. Αφαίρεση μη αλφαβητικών όρων
5. Αφαίρεση όλων των όρων με μήκος 1
6. Αφαίρεση stop-words καθώς δεν προσφέρουν πληροφορία



Προεπεξεργασία

Μετά την ολοκλήρωση της διαδικασίας προεπεξεργασίας τα δεδομένα έχουν την εξής μορφή :

```
l > read.csv("trainSet.csv")
      ,text,label
0,house dem aide even see comey letter jason chaffetz tweeted darrell lucus october subscribe jason chaffetz s
1,ever get feeling life circles roundabout rather heads straight line toward intended destination hillary clin
```



Word2Vec

Η μέθοδος Word2Vec είναι ένα σύνολο αρχιτεκτονικών μοντέλων για text vectorization και χρησιμοποιούνται για την δημιουργία Word embeddings. Υπάρχουν δύο κύριοι αλγόριθμοι :

1. Continuous Skip-gram Model

Αυτό το μοντέλο

προβλέπει λέξεις σε ένα διάστημα - παράθυρο πριν και μετά την τρέχουσα λέξη της πρότασης. Για κάθε λέξη το μοντέλο φτιάχνει ένα pair της τρέχουσας λέξης και μιας λέξης που βρίσκεται μέσα στο διάστημα αυτό και κάποια negative samples. Ένα negative sample είναι ένα pair μεταξύ της τρέχουσας λέξης και μιας λέξης εκτός του διαστήματος. Ο αριθμός των negative samples εξαρτάται από το μέγεθος των προτάσεων, δηλαδή για μικρές προτάσεις θα έχουμε 5-15 negative samples, ενώ για μεγάλες προτάσεις θα έχουμε 2-5 negative samples

1. Continuous Bag-of-Words Model

Αυτό το μοντέλο

The wide road shimmered in the hot sun.

`tf.keras.preprocessing.sequence.skipgrams`



(wide, road)	...	(road, shimmered)	(hot, sun)	...	(the, hot)
(2, 3)	...	(3, 7)	(6, 7)	...	(1, 6)

`tf.random.log_uniform_candidate_sampler`
(negative_samples = 4)



(wide, road)
(2, 3)



(wide, sun)	(wide, hot)	(wide, temperature)	(wide, code)
(2, 7)	(2, 6)	(2, 23)	(7, 2196)

concat and add label (pos:1/neg:0)



(wide, road)	(wide, sun)	(wide, hot)	(wide, temperature)	(wide, code)
(2, 3)	(2, 7)	(2, 6)	(2, 23)	(7, 2196)
1	0	0	0	0

build context words and labels for all vocab words



Word	Context words					⇒	Labels				
2	3	7	6	23	2196	⇒	1	0	0	0	0
23	12	6	94	17	1085	⇒	1	0	0	0	0
84	784	11	68	41	453	⇒	1	0	0	0	0
							⋮				
V	45	598	1	117	43	⇒	1	0	0	0	0



Embedding

Η διαδικασία ορισμού του embedding περιλαμβάνει τον συνδυασμό του Word2Vec μοντέλου που ορίσαμε παραπάνω μαζί με τα αποτελέσματα της συνάρτησης Tokenizer().

- Το αποτέλεσμα της διαδικασίας του gensim training είναι ένα Word2Vec.txt αρχείο με την παρακάτω μορφή:

trump	0.0112	...	0.0112
won	0.4512	...	- 0.6114
state	0.6114	...	- 0.2114
clinton	0.8187	...	0.9112



Embedding

Όσον αφορά την διαδικασία του Tokenizer, στόχος μας είναι να βρούμε το λεξιλόγιο των προτάσεων και το λεξικό.

- Το αποτέλεσμα της διαδικασίας του Tokenizer είναι ένα λεξικό με την παρακάτω μορφή καθώς και το σύνολο των λέξεων :

trump	1
won	2
state	3
clinton	4



Embedding

- Με την διαδικασία `getWeights()` κάνουμε mapping μεταξύ του Word2Vec μοντέλου και του Tokenizer δημιουργώντας τον παρακάτω πίνακα για το Embedding layer, που θα αναφέρουμε αργότερα.

1	0.0112	...	0.0112
2	0.4512	...	- 0.6114
3	0.6114	...	- 0.2114
4	0.8187	...	0.9112



Αρχιτεκτονική

Η αρχιτεκτονική του μοντέλου αναγνώρισης ψευδών ειδήσεων περιλαμβάνει τα παρακάτω βασικά στοιχεία :

1. Word Embedding - Word2Vec & Tokenizer (Input Layer)

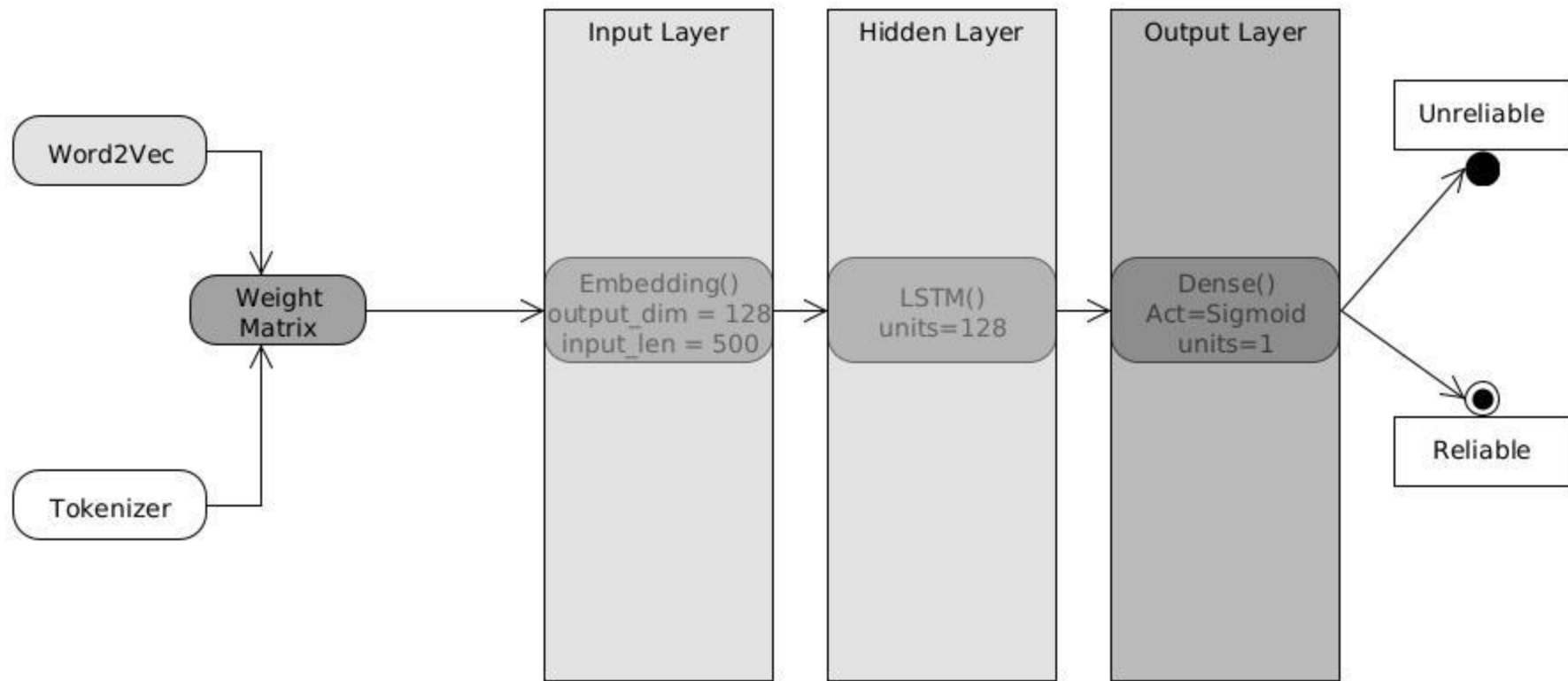
Το Word Embedding είναι μια αναπαράσταση κειμένου όπου λέξεις με την ίδια έννοια έχουν παρόμοια αναπαράσταση, καθώς οι σχετικές λέξεις που βασίζονται σε ένα κείμενο θα τοποθετούνται κοντά στο σύστημα συντεταγμένων.


2. LSTM - Long Short Term Memory (Hidden Layer)

Το LSTM είναι ικανό να θυμάται τις σημαντικές λέξεις ενός κειμένου ενώ μπορεί να "ξεχνά" τις λέξεις που δεν θεωρεί σημαντικές. Το παραπάνω γεγονός είναι ιδιαίτερα σημαντικό καθώς τα Recurrent NNs έχουν το μειονέκτημα της Short-term Memory, όπου θα δυσκολευτούν να μεταφέρουν όλη την πληροφορία από το ένα layer στο επόμενο αν το κείμενο είναι μεγάλο. Έτσι το LSTM βεβαιώνει ότι θα μεταφερθεί η σημαντική πληροφορία

3. Dense (Output Layer)

Το output layer αποτελείται από ένα Dense δίκτυο όπου λαμβάνει είσοδο από όλα τα units του LSTM, και με βάση την Sigmoid - Σιγμοειδή activation





Εκπαίδευση του Μοντέλου

Για την εκπαίδευση του μοντέλου έγιναν αρκετές προσπάθειες για την σωστή εισαγωγή παραμέτρων στο input layer, δηλαδή στο Word Embedding, Πιο συγκεκριμένα το μέγεθος του embedding είναι βέλτιστο στην τιμή 128 ενώ πολύ σημαντικό είναι και το μήκος των προτάσεων, το οποίο είναι 500, από τα αποτελέσματα του Word2Vec.

```
Epoch 1/5
97/97 [=====] - 110s 1s/step - loss: 0.5753 - accuracy: 0.6657 - val_loss: 0.4032 - val_accuracy: 0.8296
Epoch 2/5
97/97 [=====] - 111s 1s/step - loss: 0.3916 - accuracy: 0.8354 - val_loss: 0.3969 - val_accuracy: 0.8494
Epoch 3/5
97/97 [=====] - 112s 1s/step - loss: 0.3830 - accuracy: 0.8557 - val_loss: 0.3454 - val_accuracy: 0.8702
Epoch 4/5
97/97 [=====] - 111s 1s/step - loss: 0.3463 - accuracy: 0.8658 - val_loss: 0.3610 - val_accuracy: 0.8646
Epoch 5/5
97/97 [=====] - 109s 1s/step - loss: 0.3578 - accuracy: 0.8637 - val_loss: 0.3206 - val_accuracy: 0.8828
Saved model to disk
```

Αποτελέσματα Εκπαίδευσης

Αρχικά όσον αφορά το accuracy, μπορούμε να παρατηρήσουμε ότι στα validation/test δεδομένα φτάσαμε σε αποδεκτές τιμές γύρω στο 85%



Αποτελέσματα Εκπαίδευσης

Αρχικά όσον αφορά το loss, μπορούμε να παρατηρήσουμε ότι στα validation/test δεδομένα φτάσαμε ξανά σε αποδεκτές τιμές γύρω στο 30%





Ευχαριστώ για τον χρόνο σας