



## **Sentiment Analysis in Twitter**

## Περιεχόμενα

1. Εισαγωγή
2. Περιγραφή Συστήματος
3. Εκτέλεση
4. Βιβλιογραφία-Αναφορές

## 1. Εισαγωγή

### 1.1 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος, γνωστή και ως εξόρυξη γνώμης, είναι μία τεχνική μηχανικής μάθησης, η οποία αφορά την χρήση επεξεργασίας φυσικής γλώσσας, της ανάλυσης κειμένου και της υπολογιστικής γλωσσολογίας, με σκοπό την συστηματική αναγνώριση και μελέτη υποκειμενικών πληροφοριών.

Γενικά, η ανάλυση συναισθήματος στοχεύει στον προσδιορισμό της στάσης του ομιλητή-συγγραφέα σε σχέση με κάποιο θέμα. Συνήθως, γίνεται χρήση ενός δείκτη αντικειμενικότητας ή πολικότητας.

Στα πλαίσια της εργασίας θα γίνει ανάλυση συναισθήματος στο twitter.

### 1.2 Twitter

Το twitter είναι ένα κοινωνικό δίκτυο, στο οποίο οι χρήστες επικοινωνούν και αλληλεπιδρούν μεταξύ τους μέσω των tweets, τα οποία είναι μικρά κείμενα περιορισμένα σε 140 χαρακτήρες το πολύ. Επίσης, το twitter παρέχει δυνατότητα χρήσης των δεδομένων του για προγραμματιστές.

## 2. Περιγραφή Συστήματος

### 2.1 Εργαλεία και Βιβλιοθήκες

Αρχικά, η υλοποίηση της εργασίας καθώς και η γραφική διεπαφή χρήστη αναπτύχθηκε σε Java JDK 8 με Netbeans.

Για τη σύνδεση στο twitter και την εξόρυξη δεδομένων από αυτό έγινε χρήση του twitter4j, το οποίο αποτελεί την προγραμματιστική διεπαφή μεταξύ twitter και java.

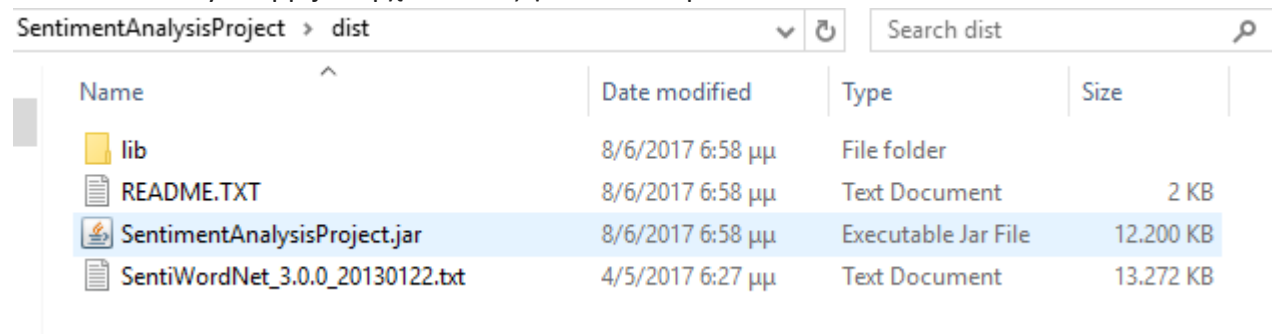
Επίσης, χρησιμοποιήθηκε η βιβλιοθήκη Google Guava για τη χρήση των δομών HashMap και HashBiMap.

Τέλος, έγινε χρήση του Stanford POS tagger για να αναγνωριστούν τα μέρη του λόγου (Parts of speech) μέσα στο κείμενο που αναλύεται.

## 2.2 SentiWordNet

Η ανάλυση του κειμένου γίνεται με χρήση του SentiWordNet, το οποίο είναι ένας λεκτικός πόρος για εξόρυξη γνώμης. Το SentiWordNet αποτελείται από συλλογές συνωνύμων που αντιστοιχίζονται σε κάποιο ID, το οποίο λειτουργεί ως δείκτης.

Για τη σωστή λειτουργία του προγράμματος πρέπει να περιλαμβάνεται το αρχείο "SentiWordNet\_3.0.0\_20130122.txt" στον φάκελο "dist" που βρίσκεται το SentimentAnalysisApp.jar αρχείο όπως φαίνεται παρακάτω



Name	Date modified	Type	Size
lib	8/6/2017 6:58 μμ	File folder	
README.TXT	8/6/2017 6:58 μμ	Text Document	2 KB
SentimentAnalysisProject.jar	8/6/2017 6:58 μμ	Executable Jar File	12.200 KB
SentiWordNet_3.0.0_20130122.txt	4/5/2017 6:27 μμ	Text Document	13.272 KB

## 2.3 Είσοδος στο twitter και Σκανάρισμα Sentiwordnet

Το σύστημα αποτελείται από 6 κλάσεις οι οποίες αλληλεπιδρούν μεταξύ τους.

- 1) MainWindow.java
- 2) Options.java
- 3) Utilities.java
- 4) AnalyzeText.java
- 5) AnalyzeTweets.java
- 6) AnalyzeUser.java

Αρχικά, όταν εκτελείται το πρόγραμμα ανοίγει το παράθυρο "MainWindow" και ζητείται από το χρήστη να εισάγει τις μεταβλητές Consumer Key, Consumer Secret, Access Token και Access Token Secret για τη σύνδεση στο twitter. Ως default υπάρχουν οι τιμές του λογαριασμού μου.

Μόλις εισαχθούν, διαβάζεται το SentiWordNet με την μέθοδο "ScanWordNet()" της κλάσης Utilities και δημιουργούνται 5 ευρετήρια ως εξής

Όνομα ευρετηρίου	Δομή δεδομένων	Παράμετροι	Παράδειγμα
partOfSpeech	HashMap	<Integer,String>	00036998, a
posScore	HashMap	<Integer,Float>	00036998, 0
negScore	HashMap	<Integer,Float>	00036998,0.5
termDictionary	HashBiMap	<Integer,List<String>>	00036998,[sluggish#2,slow#6,dull#8]

invTermDictionary	HashBiMap	<List<String>,Integer>	[sluggish#2,slow#6,dull#8],00036998
-------------------	-----------	------------------------	-------------------------------------

Χρησιμοποιώ HashMap για να είναι γρηγορότερη η αναζήτηση των πληροφοριών του κάθε σετ συνονύμων, και HashBiMaps για να δημιουργήσω το αντεστραμμένο ευρετήριο invTermDictionary, στο οποίο γίνεται η αναζήτηση των όρων κειμένου.

## 2.4 Ανάλυση κειμένου και Κατηγοριοποίηση

Αφού δημιουργηθούν τα παραπάνω ευρετήρια ανοίγει το παράθυρο “Options” το οποίο δίνει τη δυνατότητα στο χρήστη να διαλέξει μία από τις παρακάτω επιλογές :

1. Analyze tweets : Ανάλυση των πιο πρόσφατων tweets με βάση το home timeline του twitter
2. Analyze text : Ανάλυση κειμένου το οποίο εισάγεται από το χρήστη
3. Analyze user : Ανάλυση χρήστη με βάση τα πιο πρόσφατα tweets που ανέβασε στον λογαριασμό του

Και οι τρεις επιλογές λειτουργούν με την παρακάτω ακολουθία ενεργειών.

### Βήμα 1 : AnalyzeText()

1. Δεχόμαστε ως είσοδο κάποιο κείμενο (είτε από tweet είτε από τον χρήστη)
2. Χωρίζουμε το κείμενο σε tokens
3. Στο κάθε token αποδίδεται, μέσω του Stanford POS tagger, μία τιμή για το μέρος του λόγου (Part of Speech) που αποτελεί. Η τιμή αυτή βασίζεται στο Penn Treebank δέντρο.
4. Μετατρέπονται τα μέρη του λόγου σε ρήματα (verb), σε επιρρήματα (adverb), σε ουσιαστικά (noun) και σε επίθετα (adjective) σε ενικό με lemmatization\*, ώστε να ταυριάζουν με την μορφή του SentiWordNet.
5. Γίνεται αναζήτηση για όλα τα token, με την αντίστοιχη τιμή part of speech, στο αντεστραμμένο ευρετήριο όρων (invTermDictionary) και επιστρέφονται τα id τους

\*Επίσης επειδή κάποιες λέξεις δεν αποτελούν ρήμα, επίρρημα, ουσιαστικό ή επίθετο μέσα στην πρόταση, για παράδειγμα οι λέξεις and, for, but κλπ οι οποίες είναι συνδετικές λέξεις, τις παραβλέπουμε. Αυτό το κάνουμε διότι δεν υπάρχουν στο SentiWordNet και επομένως δεν μπορούν να προσθέσουν κάποια χρήσιμη πληροφορία στο συναίσθημα του κειμένου που αναλύεται.

### Βήμα 2 : CalculateScore()

1. Για κάθε id που βρέθηκε, υπολογίζουμε το σκορ  $swnScore(i)$  που του αντιστοιχεί με το άθροισμα 
$$swnScore(i) = \frac{\sum_{k=1}^j 1 + posScore(i,k) - negScore(i,k)}{j}$$
 Όπου i είναι το token του κειμένου, j είναι ο αριθμός εμφανίσεων του token στο SentiWordNet και k είναι τα id που αντιστοιχούν στο token i

2. Υπολογίζουμε το συνολικό σκορ του κειμένου από την μέση τιμή των  $swnScore(i)$ , δηλαδή

$$tweetScore = \frac{\sum_{i=1}^n swnScore(i)}{n}$$

**Σημείωση :** Επειδή ορισμένοι όροι δεν βρίσκονται στο αντεστραμμένο ευρετήριο όρων, χρησιμοποιώ το μέτρο accuracy, το οποίο είναι το ποσοστό των όρων (ρήματα, επιρρήματα, ουσιαστικά και επίθετα) του κειμένου που βρέθηκαν στο SentiWordNet.

### Βήμα 3 : ClassifyScore()

Η ανάλυση τελειώνει με την κατηγοριοποίηση του τελικού σκορ όπως φαίνεται παρακάτω :

Κατηγορία	Τιμή tweetScore
Positive	$tweetScore > 1.2$
Somewhat Positive	$0.95 < tweetScore \leq 1.2$
Neutral	$0.5 < tweetScore \leq 0.95$
Somewhat Negative	$0.2 < tweetScore \leq 0.5$
Negative	$0 \leq tweetScore \leq 0.2$
null	$tweetScore < 0$

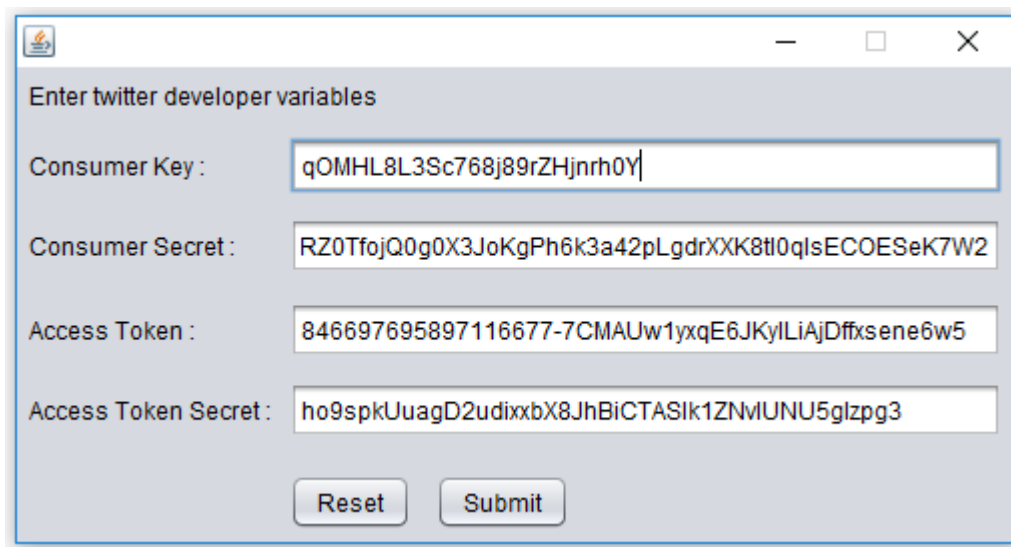
Το tweetScore μπορεί να πάρει την τιμή -1 σε περίπτωση που δεν βρεθεί κανένα token, δηλαδή έχουμε accuracy=0%

### 3. Εκτέλεση

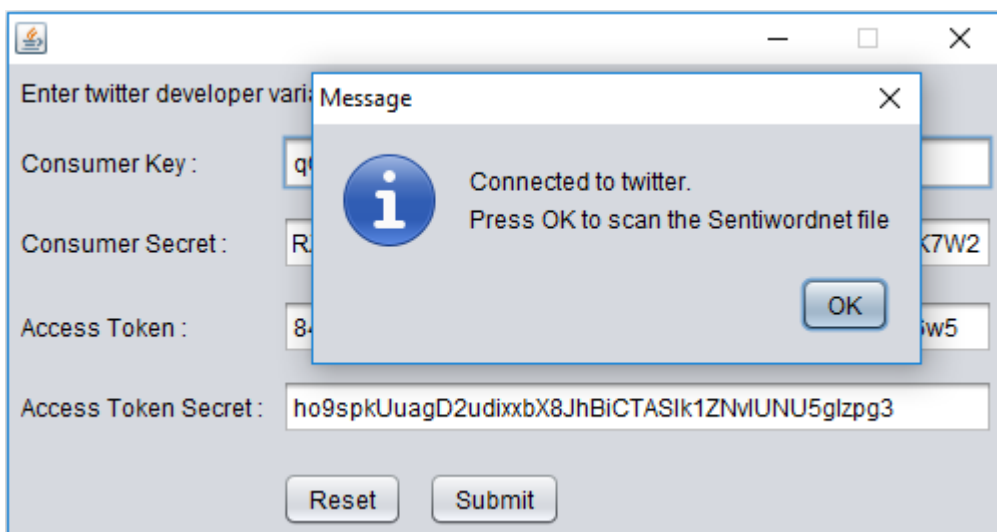
Αρχικά, εκτελούμε το αρχείο "SentimentAnalysisProject.jar" που βρίσκεται στον φάκελο "dist"

SentimentAnalysisProject > dist		Search dist	
Name	Date modified	Type	Size
lib	8/6/2017 6:58 μμ	File folder	
README.TXT	8/6/2017 6:58 μμ	Text Document	2 KB
SentimentAnalysisProject.jar	8/6/2017 6:58 μμ	Executable Jar File	12.200 KB
SentiWordNet_3.0.0_20130122.txt	4/5/2017 6:27 μμ	Text Document	13.272 KB

Ανοίγει το παράθυρο και μας ζητούνται τα κλειδιά για την είσοδο στο twitter.Εισάγουμε τα κλειδιά και πατάμε το κουμπί Submit



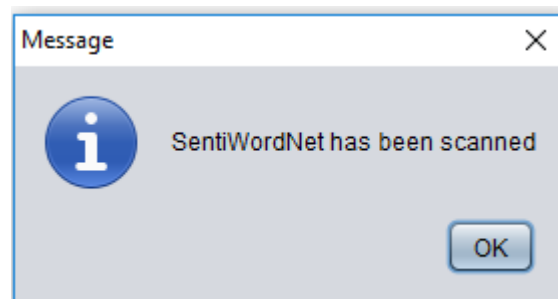
A screenshot of a Windows-style dialog box titled "Enter twitter developer variables". It contains four text input fields with the following values: "Consumer Key : qOMHL8L3Sc768j89rZHjnrh0Y", "Consumer Secret : RZ0TfojQ0g0X3JoKgPh6k3a42pLgdrXXK8tl0qlsECOESek7W2", "Access Token : 846697695897116677-7CMAUw1yxqE6JKyLiAjDffxsene6w5", and "Access Token Secret : ho9spkUuagD2udixbX8JhBiCTASIk1ZNMUNU5glzpg3". At the bottom are "Reset" and "Submit" buttons.



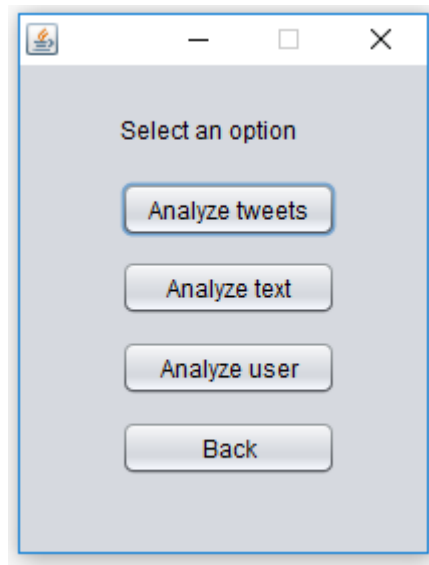
A screenshot of the same dialog box as above, but with a smaller "Message" dialog box overlaid in the center. The message box has an information icon and contains the text: "Connected to twitter. Press OK to scan the Sentiwordnet file". An "OK" button is at the bottom right of the message box.

Πατάμε στο κουμπί "OK" και περιμένουμε να διαβαστεί το αρχείο SentiWordNet και να αποθηκευτούν οι κατάλληλες πληροφορίες από αυτό.

Μόλις ολοκληρωθεί,εμφανίζεται το παρακάτω μήνυμα,πατάμε το κουμπί "OK" και ανοίγει το παράθυρο των επιλογών

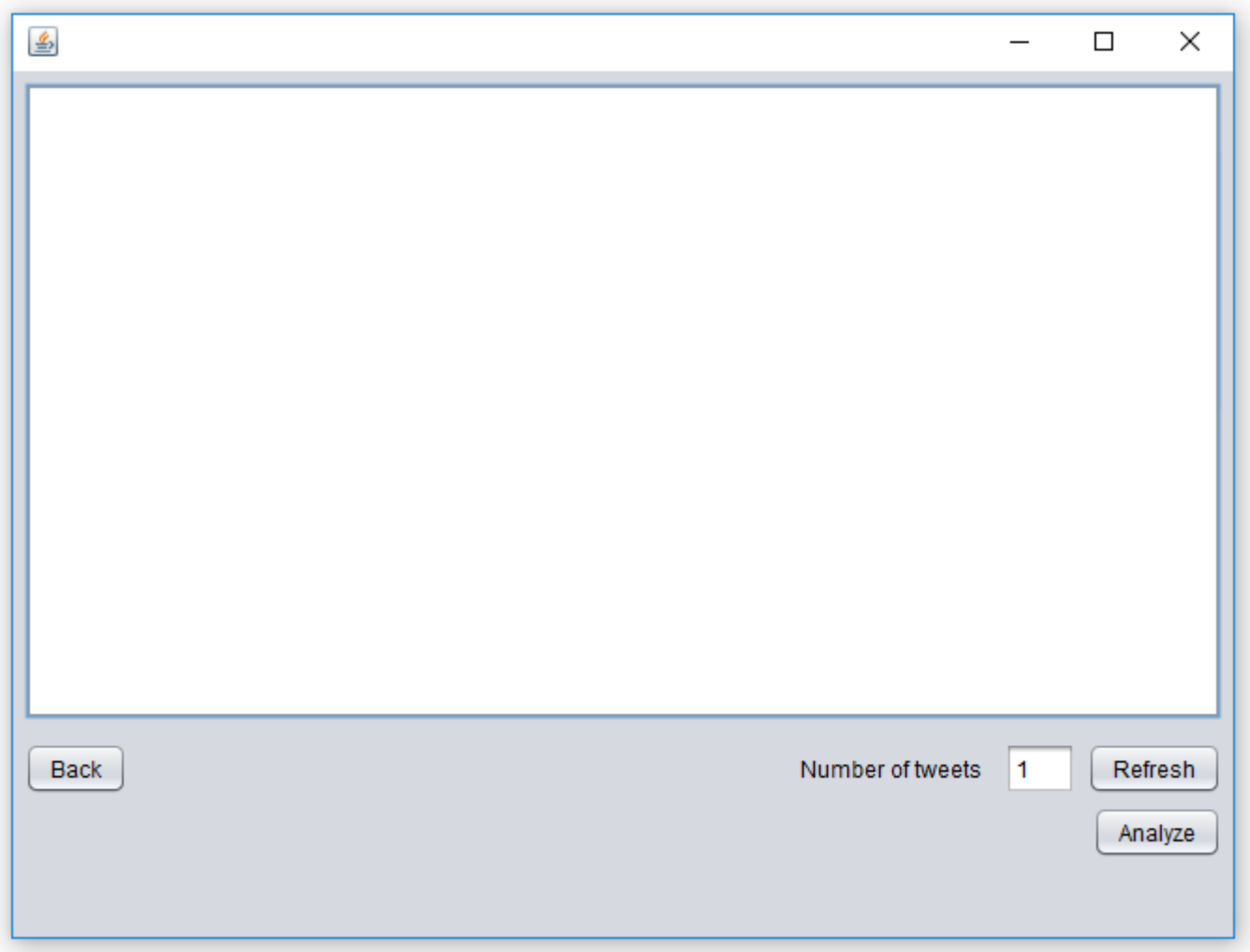


A screenshot of a "Message" dialog box with an information icon and the text: "SentiWordNet has been scanned". An "OK" button is at the bottom right.

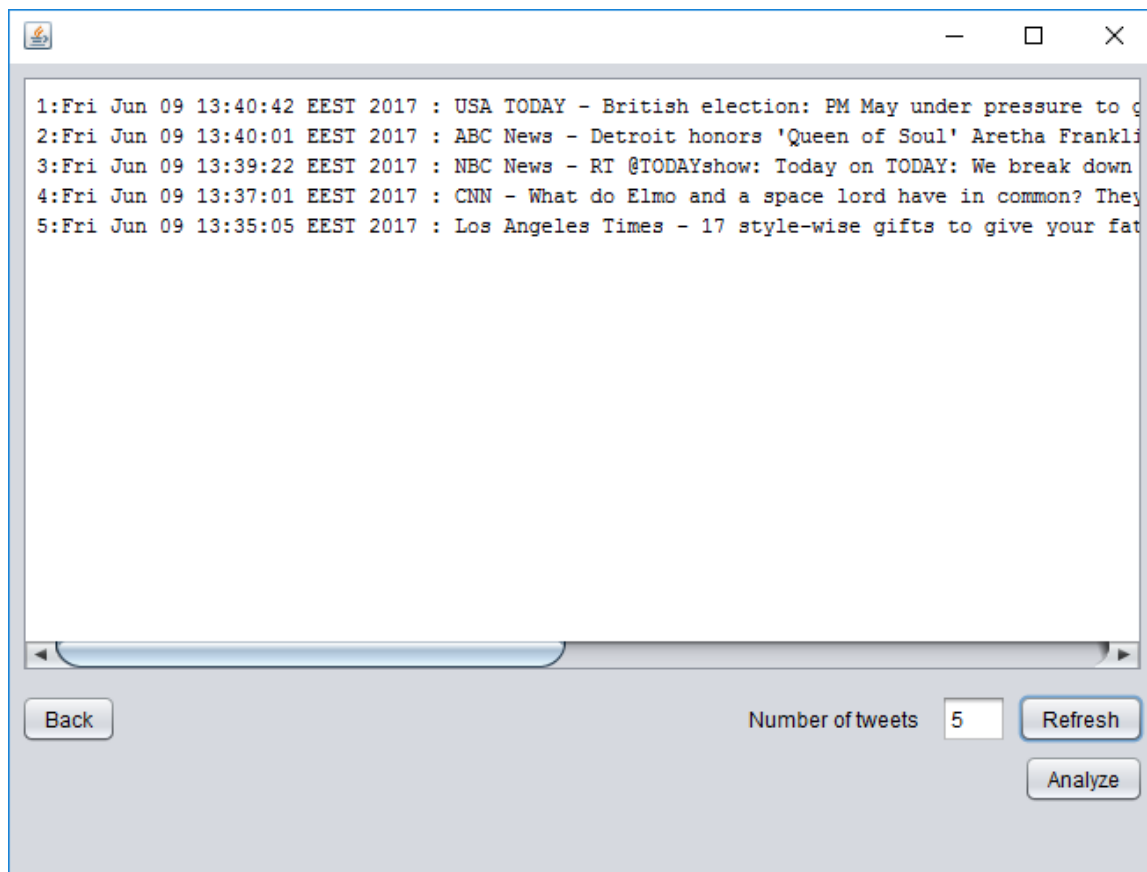


### Επιλογή “Analyze tweets”

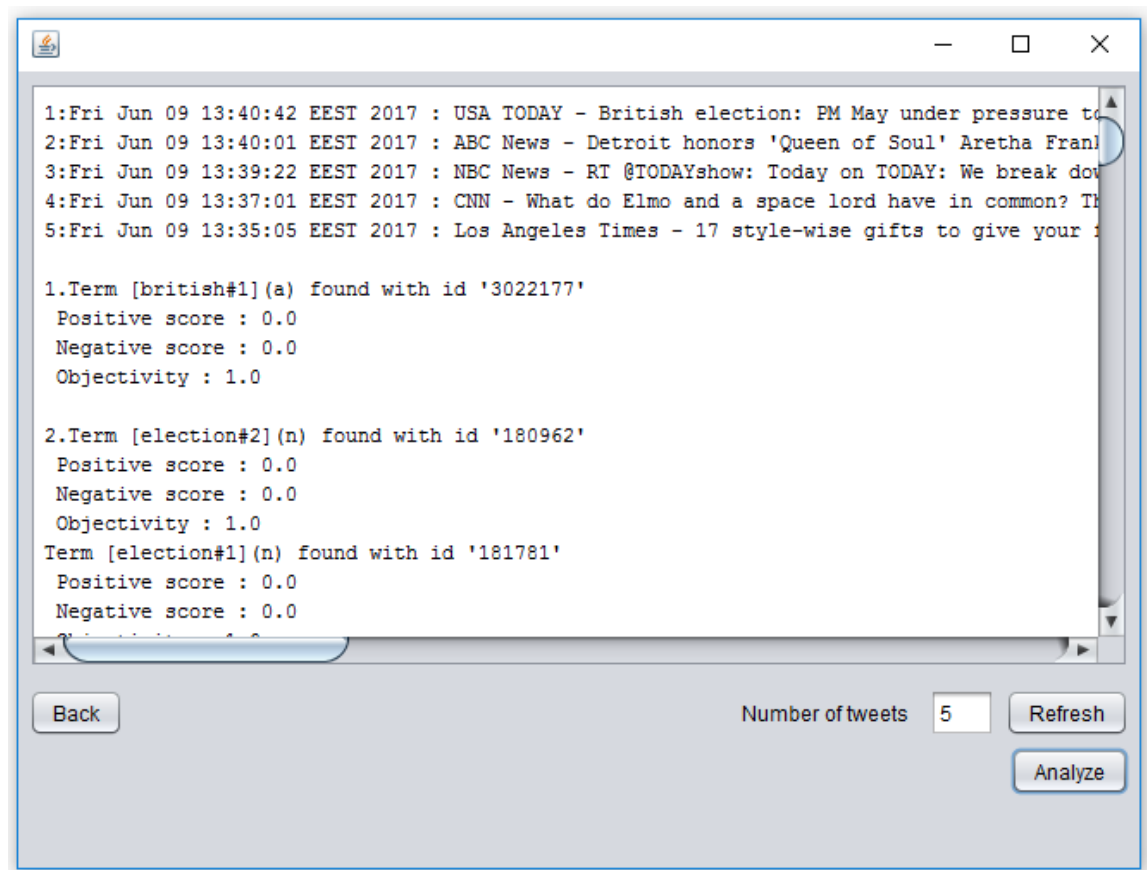
Στο παράθυρο αυτό γίνεται αναζήτηση των πιο πρόσφατων tweets από το home timeline.



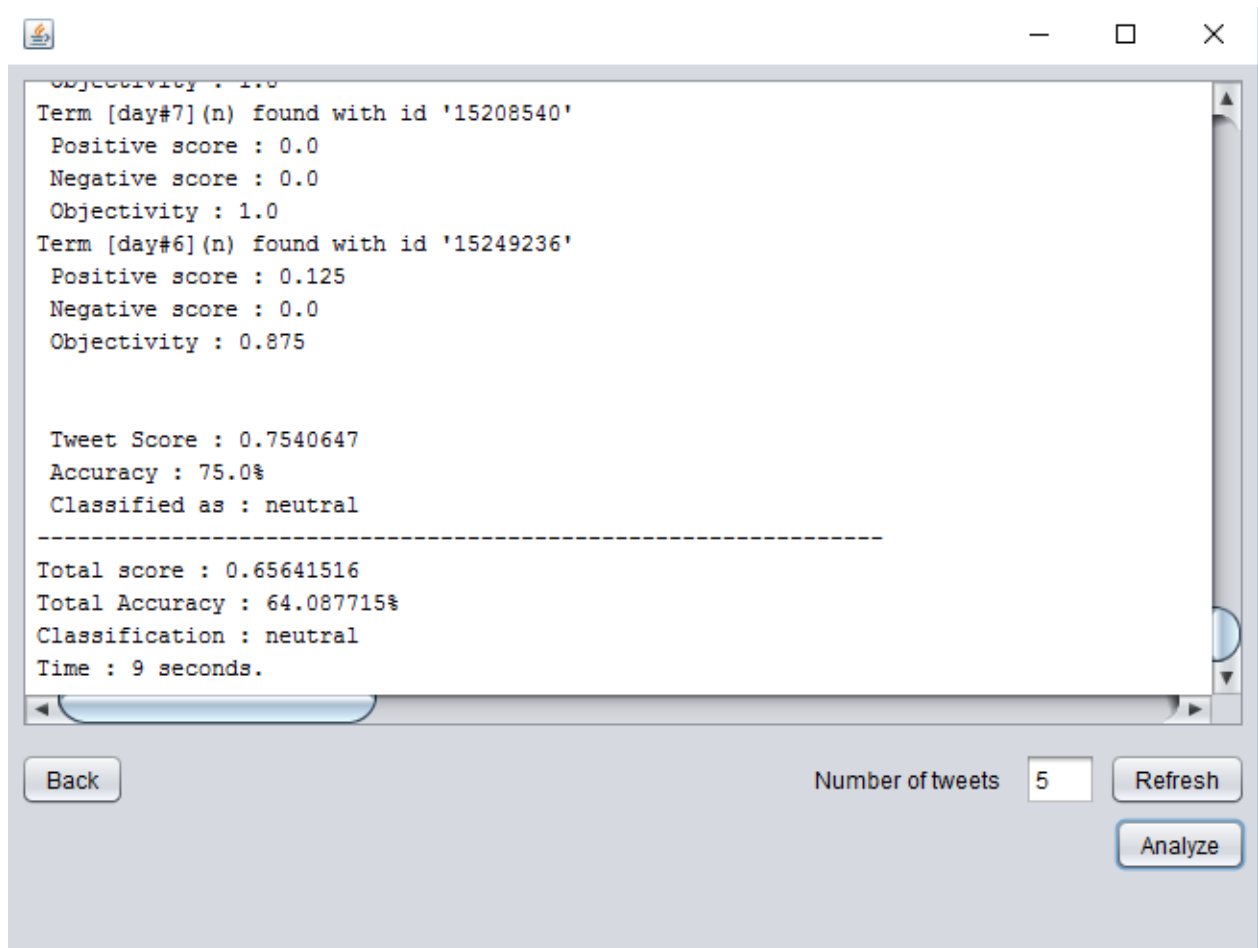
Καταρχάς, εισάγουμε τον αριθμό των tweets που θέλουμε να αναλυθούν στο πεδίο “Number of tweets” και πατάμε το κουμπί “Refresh” για να εμφανιστούν



Για να αναλύσουμε τα tweets πατάμε το κουμπί “Analyze” και μας εμφανίζονται τα ακόλουθα αποτελέσματα







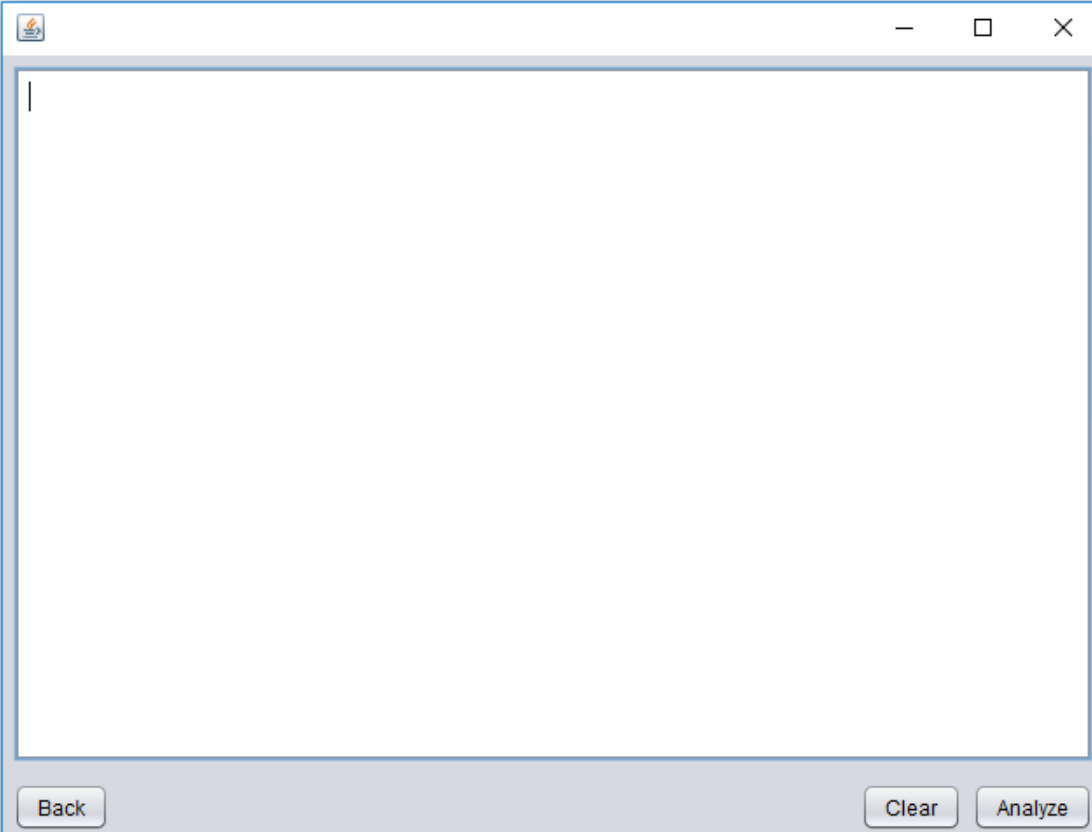
Αρχικά εμφανίζεται κάθε όρος που βρέθηκε στο SentiWordNet, μαζί με το μέρος του λόγου του, το id που του αντιστοιχεί, το θετικό και το αρνητικό σκορ του, και το μέτρο αντικειμενικότητας των βαθμολογιών του. Η αντικειμενικότητα βαθμολογίας κάθε όρου υπολογίζεται ως  $obj(i) = 1 - (posScore(i) + negScore(i))$

Αφού αναζητηθούν όλοι οι όροι εμφανίζεται το σκορ του tweet (Tweet Score), η ακρίβεια (accuracy) και η κατηγοριοποίηση του tweet (Classified as).

Στο τέλος εμφανίζεται το μέσο σκορ των tweets (Total score), η μέση ακρίβεια (Total accuracy) των tweets, η μέση κατηγοριοποίηση (Classification) και ο συνολικός χρόνος που χρειάστηκε για την εκτέλεση των λειτουργιών.

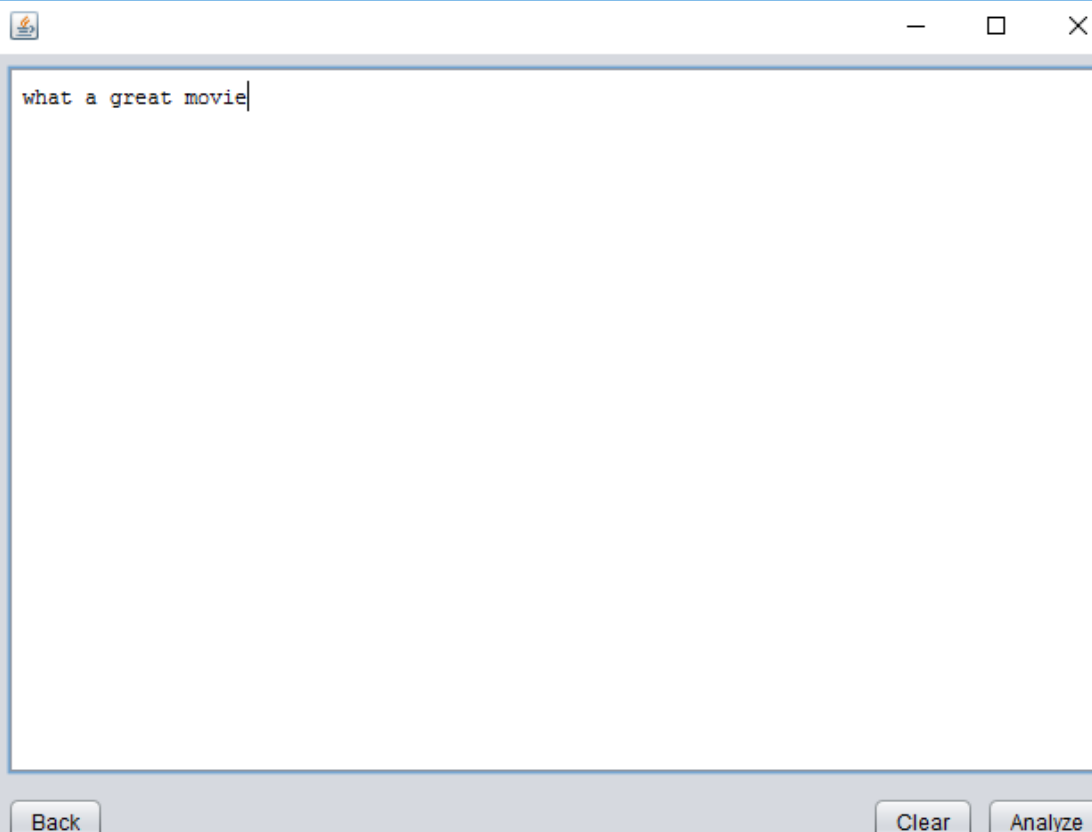
## Επιλογή “Analyze text”

Στο παράθυρο αυτό η ανάλυση γίνεται πάνω σε κείμενο που εισάγει ο χρήστης.



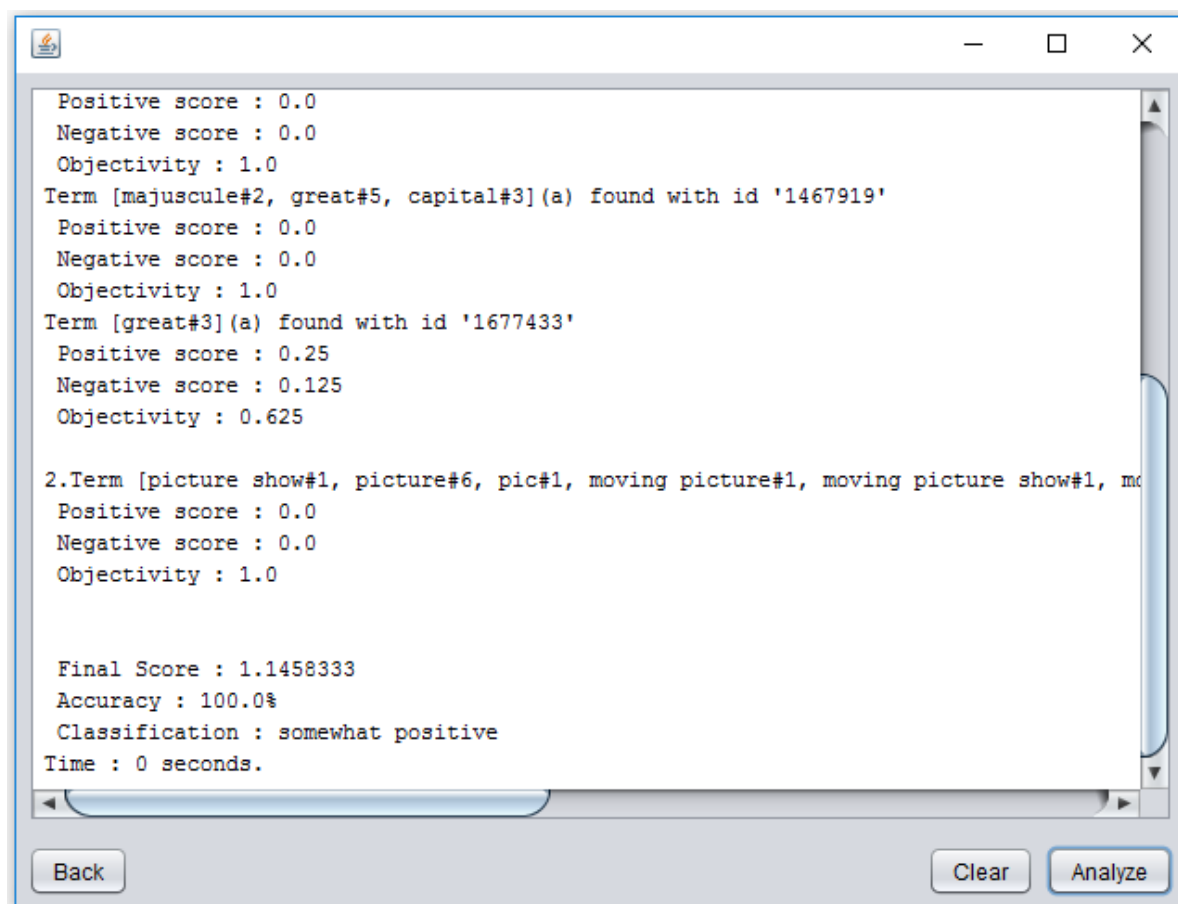
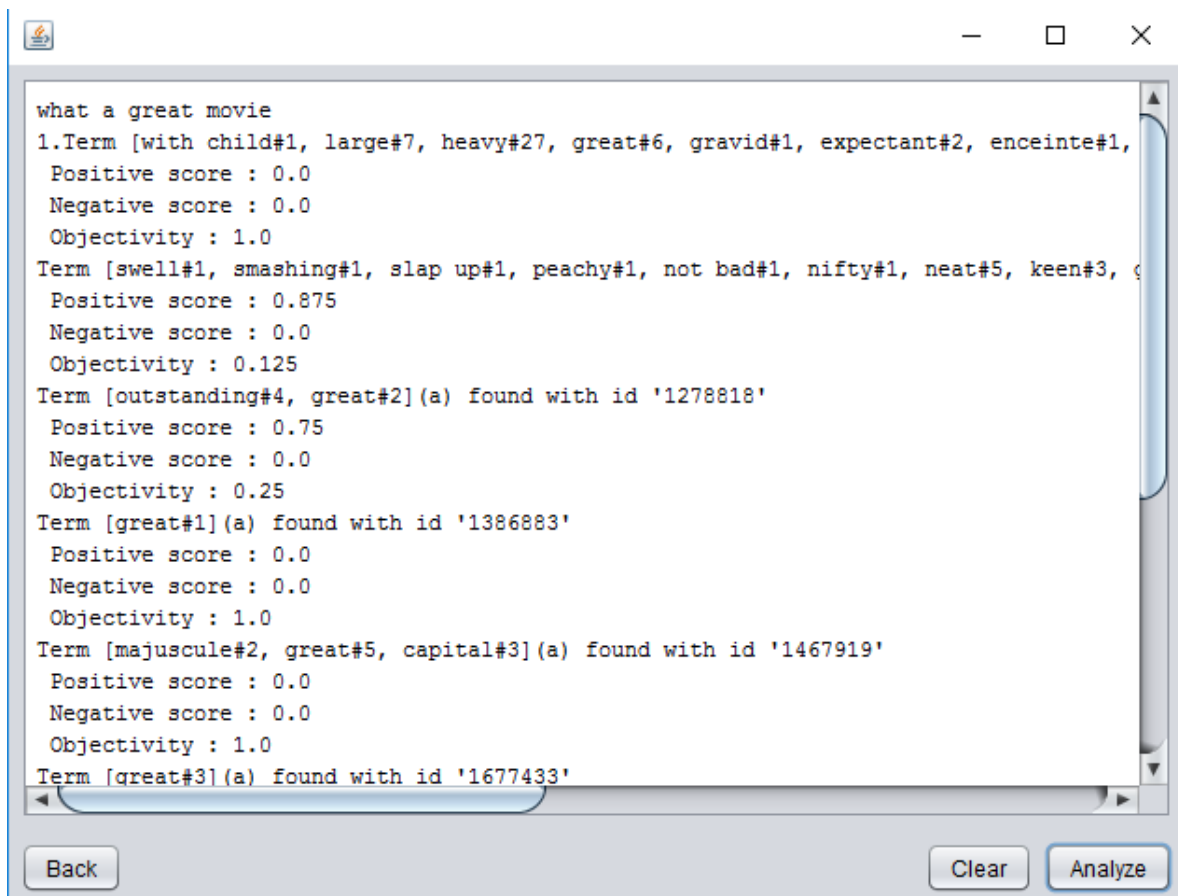
A screenshot of a software window titled "Analyze text". The window has a standard Windows-style title bar with minimize, maximize, and close buttons. The main area is a large, empty text input field with a vertical cursor at the top left. At the bottom of the window, there is a light gray bar containing three buttons: "Back" on the left, and "Clear" and "Analyze" on the right.

Αρχικά εισάγουμε το κείμενο που θέλουμε να αναλυθεί



A screenshot of the same "Analyze text" window. The text input field now contains the text "what a great movie" in a monospaced font, with a vertical cursor at the end of the text. The "Back", "Clear", and "Analyze" buttons remain at the bottom.

Πατάμε το κουμπί “Analyze” και μας εμφανίζονται τα ακόλουθα

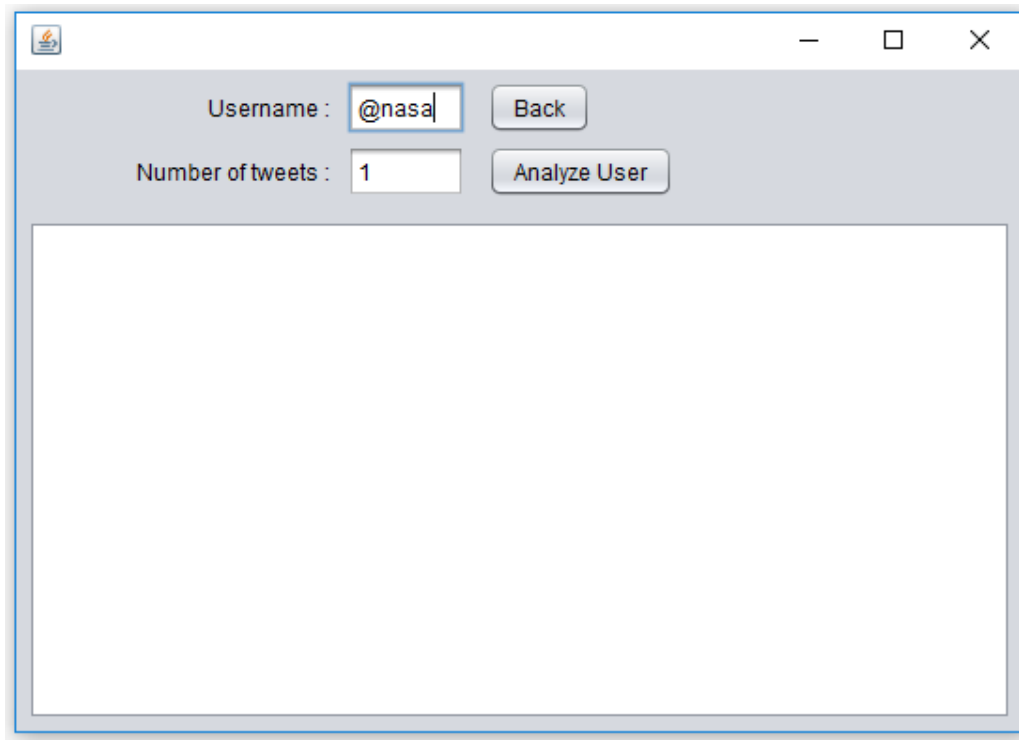


Όπως και στο Analyze Tweets, στο τέλος μας εμφανίζονται το τελικό σκορ(Final Score), η ακρίβεια(Accuracy), η κατηγοριοποίηση(Classification) και ο χρόνος εκτέλεσης(Time).

Για να σβήσουμε τα αποτελέσματα πατάμε το κουμπί “Clear”.

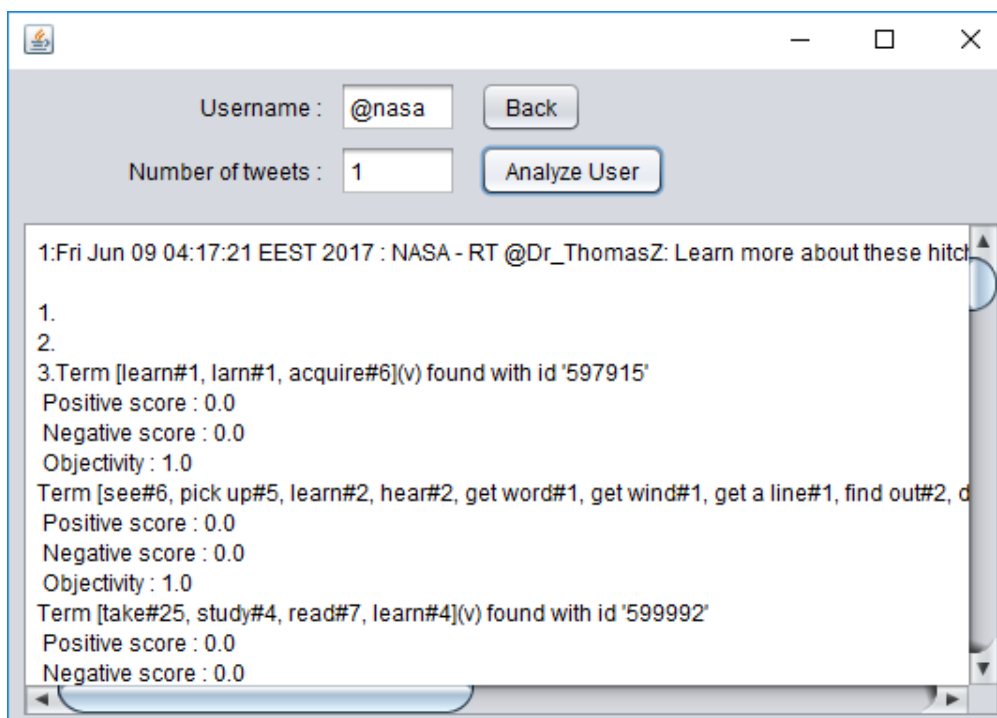
### Επιλογή “Analyze user”

Η επιλογή Analyze User κάνει ανάλυση στα πιο πρόσφατα tweets από το timeline ενός χρήστη του twitter

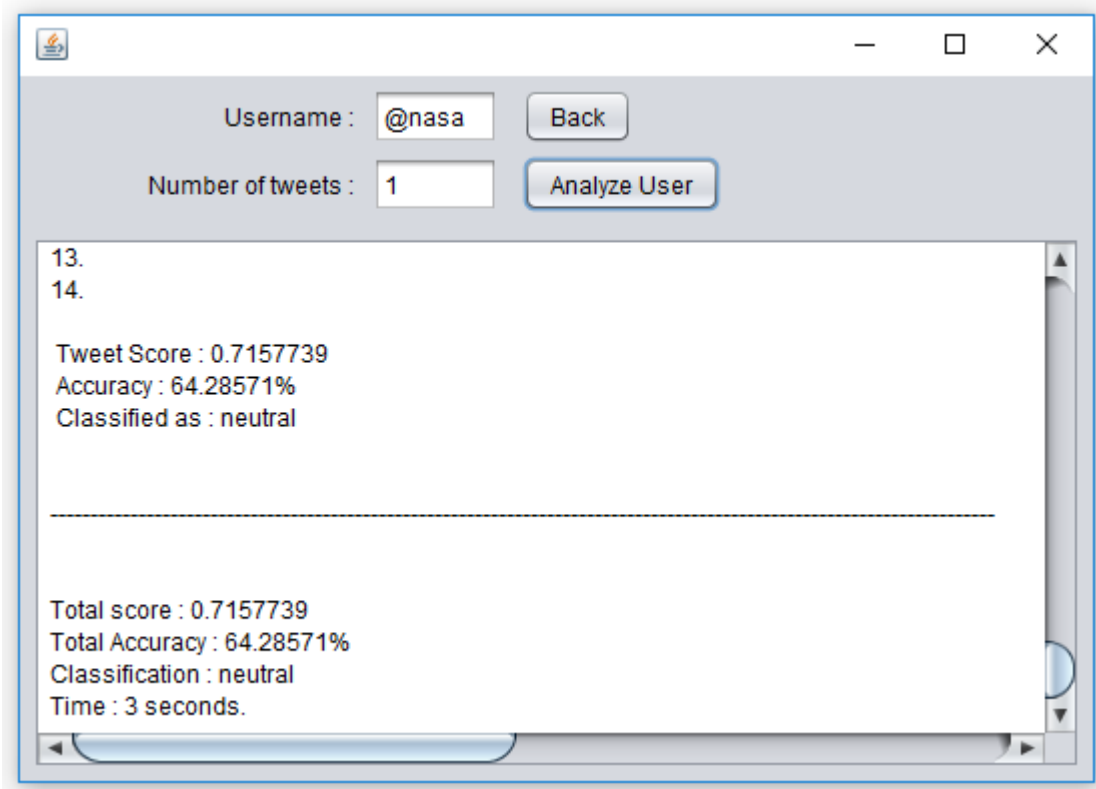


The screenshot shows a software window titled "Analyze User". It has a light gray background. At the top left is a small icon of a notepad. The window contains two input fields: "Username :" with the text "@nasa" and "Number of tweets :" with the value "1". To the right of the first field is a "Back" button. To the right of the second field is an "Analyze User" button. Below these fields is a large, empty rectangular area for displaying results.

Αρχικά εισάγουμε τον χρήστη που θέλουμε να αναλύσουμε και τον αριθμό των tweets,πατάμε το κουμπί “Analyze User” και εμφανίζονται τα αποτελέσματα



The screenshot shows the same "Analyze User" window, but now the large rectangular area contains text results. The text is as follows:  
1:Fri Jun 09 04:17:21 EEST 2017 : NASA - RT @Dr\_ThomasZ: Learn more about these hitc  
1.  
2.  
3.Term [learn#1, larn#1, acquire#6](v) found with id '597915'  
Positive score : 0.0  
Negative score : 0.0  
Objectivity : 1.0  
Term [see#6, pick up#5, learn#2, hear#2, get word#1, get wind#1, get a line#1, find out#2, d  
Positive score : 0.0  
Negative score : 0.0  
Objectivity : 1.0  
Term [take#25, study#4, read#7, learn#4](v) found with id '599992'  
Positive score : 0.0  
Negative score : 0.0



Όμοια με πριν, εμφανίζονται οι όροι των tweet που βρέθηκαν στο SentiWordNet και τα μεγέθη Tweet Score, Accuracy και Classified as για το κάθε tweet. Στο τέλος, υπολογίζονται και εμφανίζονται τα χαρακτηριστικά Total Score, Total Accuracy, Classification και Time

#### 4. Βιβλιογραφία-Αναφορές

[1] Twitter4j api

<http://twitter4j.org/en/>

[2] SentiWordNet

<http://sentiwordnet.isti.cnr.it/>

[3] Parts of speech

[http://www.academia.edu/4062253/Using\\_SentiWordNet\\_for\\_Sentiment\\_Classification\\_What\\_is\\_SentiWordNet](http://www.academia.edu/4062253/Using_SentiWordNet_for_Sentiment_Classification_What_is_SentiWordNet)

[4] Paging

<http://stackoverflow.com/questions/24838833/twitter-gethometimeline-always-return-me-only-20-status>

[5] Regex

<https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html>

[6] Regex testing

<http://www.regexpplanet.com/advanced/java/index.html>

[7] SentiWordNet demo code

<http://sentiwordnet.isti.cnr.it/code/SentiWordNetDemoCode.java>

[8] Map

<https://docs.oracle.com/javase/8/docs/api/java/util/Map.html>

[9] List

<https://docs.oracle.com/javase/7/docs/api/java/util/List.html>

[10] Google Guava library

<https://github.com/google/guava/wiki/Release19>

[11] Iterate over map entries

<http://stackoverflow.com/questions/46898/how-to-efficiently-iterate-over-each-entry-in-a-map>

[12] Classification

<http://www.ds.unipi.gr/prof/cdoulik/papers/semEval15.pdf>

[13] Stanford POS tagger

<https://nlp.stanford.edu/software/tagger.shtml>

[14] Penn Treebank POS tags

[https://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

[15] Stanford POS tagger tutorial

<http://new.galalaly.me/index.php/2011/05/tagging-text-with-stanford-pos-tagger-in-java-applications/>