# Progress File

Chi squere

Δουλεύει με tf-idf weighting το πήρα έτομο από το sklearn

Χρησιμοποιεί το selectkbest όπως υπάρχει στο sklearn example

RDF

Χρησιμοποιεί το countvectorizer για να κάνει ένα αρχικό relative\_term pool  
Φτιάχνει ένα αρχικό train\_set  
Στην συνέχεια φτιάχνει μια λίστα με τους υποψήφιους προς αποχώρηση όρους και τους αφαιρεί από την αρχικό train\_set .  
Κατόπιν παίρνει το νέο train\_set και το κάνει tf-idf. ?? ( ρώτα αραμπατζή)  
Για να μειώσω τα χαρακτηριστικά βάζω threshold min\_df= 12 δηλαδή θα πρέπει ο όρος να εμφανίζεται τουλάχιστον σε 12 κείμενα relative.

Example\_unifromity\_Cumulative

Χρησιμοποιεί countvectorizer για να πάρει τα relative feat   
Φτιάχνει το timeline το οποίο είναι ουσιαστικά το άθροισμα των relative κειμένων του train\_set  
Το step= 1/ timeline  
Υπόθεση αν ένα χαρακτηριστικό είναι uniform τότε θα εμφανίζεται τουλάχιστον από μια φορά σε κάθε relative document του train\_set  
Οπότε φτιάχνουμε την βέλτιστη συναθροιστική uniform κατανομή η οποία είναι ίσα steps από 0 εώς 1  
Φτιάχνουμε για κάθε feature την δικιά του κατανομή με το εξής κριτήριο :

* Αν έχει occurrence παίρνει κ\*βήμα
* Αν δεν έχει μένει στάσιμο
* Στο τέλος έχει 1

Έπειτα χρησιμοποιώ επαναληπτικά την ks2samp από scipy.stats και παίρνω τα pvalues για κάθε feature.  
Αποθηκεύω τα ζευγάρια p\_value – feature σε νέο dataframe και μετά το κάνω sort . Από εκεί μπορώ να πάρω τους τοπ όρους που θέλω.

FeatureSelection class