ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ.
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ.

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ - ΓΙΩΡΓΟΣ ΛΙΑΣΣΑΣ**

# ΑΥΤΟΜΑΤΗ ΠΑΡΑΓΩΓΗ ΠΕΡΙΛΗΨΕΩΝ ΑΠΟ ΙΣΤΟΛΟΓΙΑ

**Επιβλέπων Καθηγητής**

**ΙΩΝ ΑΝΔΡΟΥΤΣΟΠΟΥΛΟΣ**

**Βοηθός Επίβλεψης**

**ΔΗΜΗΤΡΗΣ ΓΑΛΑΝΗΣ**

**Αθήνα, Νοέμβριος 2010**

# ΠΕΡΙΕΧΟΜΕΝΑ

**1. Εισαγωγή .....2**

1.1 Αντικείμενο της εργασίας..... 2

1.2 Διάρθρωση της εργασίας ..... 2

**2. Συστήματα Παραγωγής περιλήψεων .....3**

2.1 Είδη περιλήψεων..... 3

2.2 Βασικά στάδια παραγωγής περίληψης ..... 4

2.3 Προγενέστερα συστήματα παραγωγής περιλήψεων. . 5

2.3.1 Παραγωγή περιλήψεων από ειδησεογραφικά κείμενα ..... 5

2.3.2 Παραγωγή περιλήψεων από ιστολόγια..... 11

**3. Το σύστημα και τα πειράματα της εργασίας .....14**

3.1 Προεπεξεργασία ιστολογίων. ..... 15

3.2 Πειράματα ..... 16

Πείραμα 1. ..... 18

Πείραμα 2. ..... 21

Παρατηρήσεις..... 24

**4. Συμπεράσματα και μελλοντικές κατευθύνσεις.....26**

**Βιβλιογραφία .....27**

# 1. Εισαγωγή

## 1.1 Αντικείμενο της εργασίας

Η αύξηση των ιστολογίων (blogs) στον παγκόσμιο ιστό είναι δεδομένη. Ολοένα και περισσότεροι άνθρωποι διατηρούν το δικό τους ιστολόγιο, ενώ πολύ μεγαλύτερος είναι ο αριθμός εκείνων που διαβάζουν το περιεχόμενό τους. Η ύπαρξη ενός αποτελεσματικού συστήματος που θα μπορούσε να συνοψίσει το περιεχόμενο των ιστολογίων θα διευκόλυνε πολύ τους αναγνώστες, καθώς θα τους απάλλασσε από την χρονοβόρα ανάγνωση μεγάλου όγκου κειμένων.

Στόχος αυτής της εργασίας ήταν η δημιουργία ενός συστήματος παραγωγής περιλήψεων από ιστολόγια. Πιο συγκεκριμένα, το σύστημα που αναπτύχθηκε έχει την δυνατότητα να παράγει περιλήψεις που απαντούν σε αγγλικές ερωτήσεις του χρήστη (π.χ. «Why do people like Starbucks better than Dunkin Donuts?») βάσει ενός συνόλου από αγγλικά ιστολόγια που αναφέρονται στο ίδιο θέμα (π.χ. την αλυσίδα καταστημάτων Starbucks). Οι ερωτήσεις αυτές αναζητούν τις θετικές και αρνητικές γνώμες για κάποιο συγκεκριμένο θέμα (π.χ. ανθρώπους, προιόντα, εταιρείες, γεγονότα κ.λ.π). Τέτοιες γνώμες εκφράζονται πολύ συχνά σε ιστολόγια, ενώ πολύ σπανιότερα σε άλλου είδους δικτυακούς τόπους.

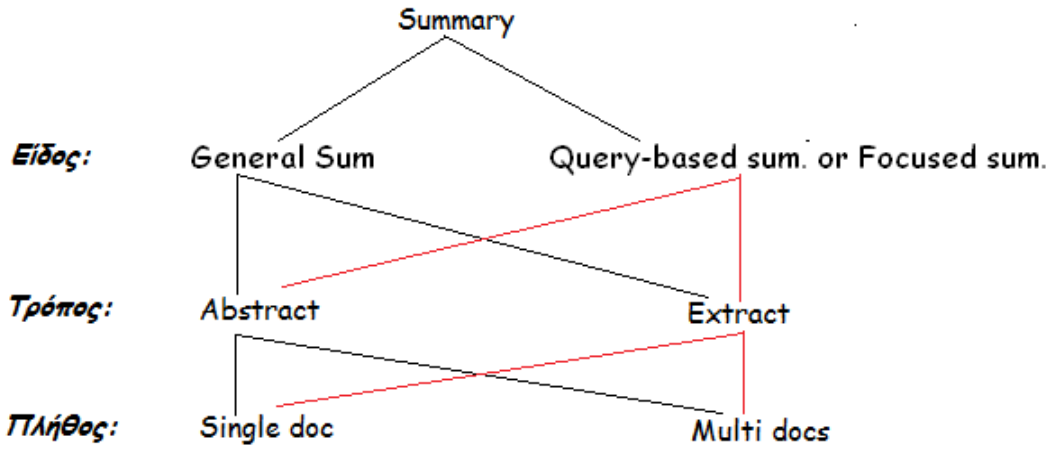
## 1.2 Διάρθρωση της εργασίας

Στο 2ο κεφάλαιο παρουσιάζονται τα βασικά στάδια των συστημάτων παραγωγής περιλήψεων και γίνεται μια συνοπτική παρουσίαση προγενέστερων συστημάτων περιλήψεων από ειδησεογραφικά κείμενα και ιστολόγια. Στο 3ο κεφάλαιο παρουσιάζεται το σύστημα της εργασίας και η πειραματική αξιολόγησή του. Τέλος, στο 4ο κεφάλαιο γίνεται μια ανασκόπηση της εργασίας, συνοψίζονται τα συμπεράσματα που προέκυψαν από τα πειράματα και προτείνονται περαιτέρω βελτιώσεις του συστήματος.

# 2. Συστήματα Παραγωγής περιλήψεων

## 2.1 Είδη περιλήψεων

Οι περιλήψεις μπορούν να κατηγοριοποιηθούν βάσει διαφόρων κριτηρίων. Το παρακάτω σχεδιάγραμμα παρουσιάζει συνοπτικά αυτά τα κριτήρια αλλά και τις μεταξύ τους σχέσεις.




Εικόνα 1 - Χαρακτηριστικά περιλήψεων και μεταξύ τους σχέσεις.

Όπως φαίνεται στην Εικόνα 1, κάθε περίληψη διακρίνεται από τρία βασικά χαρακτηριστικά. Το *είδος* της δηλώνει αν η περίληψη δημιουργήθηκε με σκοπό να απαντήσει μια ερώτηση ενός χρήστη βάσει ενός συνόλου κειμένων (**query-focused summary**) ή να συνοψίσει τα πιο σημαντικότερα στοιχεία αυτών (**general summary**). Ο *τρόπος* δηλώνει αν η περίληψη είναι αποτέλεσμα εξαγωγής προτάσεων από το σύνολο των κειμένων (**extractive summary**) ή παραγωγής νέου κειμένου χρησιμοποιώντας διαφορετικές λέξεις και διατυπώσεις για να εκφραστεί το περιεχόμενο των κειμένων (**abstractive summary**). Τέλος, το *πλήθος* δηλώνει το κατά πόσο η περίληψη παράγεται για να συνοψίσει ένα ή περισσότερα κείμενα (**multi-document or single document summary**).

## 2.2 Βασικά στάδια παραγωγής περίληψης

Παρόλο που όπως είδαμε υπάρχουν διαφορετικά είδη περιλήψεων και επομένως διαφορετικά συστήματα παραγωγής περιλήψεων, τα βασικά βήματα που ακολουθούν αυτά παραμένουν ίδια σε κάθε περίπτωση [JurMartin08]. Τα βήματα αυτά είναι τα ακόλουθα:

- ***Επιλογή περιεχομένου – (Content Selection):*** Στο πρώτο βήμα γίνεται η επιλογή του περιεχομένου που θα εισαχθεί στην περίληψη. Αυτό γίνεται από ένα σύνολο κειμένων τα οποία έχουν συνήθως επιστραφεί από ένα σύστημα ανάκτησης πληροφοριών (information retrieval). Τα κείμενα μετατρέπονται σε μια λίστα από προτάσεις, από τις οποίες επιλέγονται ορισμένες για να εισαχθούν στην περίληψη. Οι προτάσεις που θα επιλεγούν θα πρέπει να απαντούν στην ερώτηση του χρήστη (αν υπάρχει) ή να εκφράζουν τις σημαντικότερες πληροφορίες των κειμένων.
- ***Διάταξη προτάσεων – (Information Ordering):*** Το δεύτερο βήμα περιλαμβάνει τη διάταξη των προτάσεων που έχουν επιλεγεί στο προηγούμενο στάδιο. Η διάταξη των προτάσεων της περίληψης επηρεάζει την αναγνωσιμότητά της και την κατανόηση του περιεχομένου της.
- ***Πραγμάτωση προτάσεων – (Sentence Realization):*** Το τελευταίο βήμα περιλαμβάνει την αναδιατύπωση μερών των επιλεγμένων προτάσεων έτσι ώστε αυτές να είναι ευανάγνωστες και συναφείς στα πλαίσια της περίληψης.

Η αυτόματη παραγωγή περιλήψεων από πολλά κείμενα, με την οποία θα ασχοληθούμε σε αυτή την εργασία, απαιτεί την ανάπτυξη αλγορίθμων για τα προηγούμενα βήματα, αλλά παρουσιάζει μια επιπλέον δυσκολία. Σε αντίθεση με την παραγωγή περιλήψεων από ένα κείμενο, οι ίδιες πληροφορίες διατυπωμένες με παρόμοιο ή διαφορετικό τρόπο επαναλαμβάνονται σε διάφορα από τα κείμενα εισόδου. Αυτό επιδρά αρνητικά καθώς εμφανίζεται *πλεονασμός* (**redundancy**) στην παραγόμενη περίληψη. Για αυτό χρησιμοποιούνται αλγόριθμοι όπως ο MMR (**Maximal Marginal Relevance**) [Carbonell and Coldstein, 1998, Goldstein κ.ά 2000] ή αλγόριθμοι που χρησιμοποιούν *συσταδοποίηση* (**clustering**) [JurMartin08] με σκοπό τον περιορισμό του πλεονασμού. Συγκεκριμένα, ο MMR βασίζεται στη χρήση ενός *παράγοντα ποινής* (**penalty term**) για κάθε πρόταση

που είναι υποψήφια για να εισαχθεί στην περίληψη. Η τιμή του παράγοντα ποινής κάθε πρότασης εξαρτάται από την ομοιότητα αυτής με τις προτάσεις που έχουν ήδη συμπεριληφθεί στην περίληψη. Όσο μεγαλύτερη είναι η ομοιότητα τόσο μεγαλύτερες τιμές παίρνει ο παράγοντας ποινής, ο οποίος αφαιρείται από το βαθμό καταλληλότητας της πρότασης. Σε κάθε βήμα, η πρόταση με το μεγαλύτερο βαθμό καταλληλότητας επιλέγεται για να εισαχθεί στην περίληψη.

Οι αλγόριθμοι συσταδοποίησης χρησιμοποιούν μια άλλη τεχνική. Επιχειρούν να δημιουργήσουν συστάδες (ομάδες) από σχετικές προτάσεις που εκφράζουν τις ίδιες πληροφορίες και στην συνέχεια επιλέγουν την πιο αντιπροσωπευτική (**centroid**) ώστε να την περιλάβουν στην περίληψη.

## 2.3 Προγενέστερα συστήματα παραγωγής περιλήψεων.

Σε αυτή την ενότητα περιγράφονται συστήματα που παράγουν περιλήψεις είτε από ειδησεογραφικά κείμενα είτε από ιστολόγια. Τα συστήματα αυτά έχουν συμμετάσχει σε διεθνείς διαγωνισμούς και έχουν καταταγεί ανάμεσα στα κορυφαία, κυρίως σε ό,τι αφορά την επιλογή περιεχομένου. Τέτοιοι διαγωνισμοί έχουν διοργανωθεί στα πλαίσια του Document Understanding Conference (DUC, 2000-2007) και του Text Analysis Conference[1] (TAC, 2008-2010), που αποτελεί συνέχεια του DUC. [2]

### 2.3.1 Παραγωγή περιλήψεων από ειδησεογραφικά κείμενα

Τα τελευταία χρόνια (2005-2010) οι διαγωνισμοί που γίνονται στα πλαίσια των διαγωνισμών των συνεδρίων DUC και TAC αφορούν κυρίως την παραγωγή περιλήψεων από ειδησεογραφικά κείμενα. Συγκεκριμένα οι διοργανωτές δίνουν σύνολα εγγράφων και για κάθε ένα από αυτά ζητείται να παραχθεί μια περίληψη που να απαντά μια πολύπλοκη ερώτηση του χρήστη (π.χ. «Describe the state of teaching art and music in public schools around the world. Indicate problems, progress and failures.»). Παρακάτω περιγράφονται δύο συστήματα που παράγουν

[1] http://www.nist.gov/tac/
[2] http://duc.nist.gov/

τέτοιες περιλήψεις και έχουν συμμετάσχει στους διαγωνισμούς DUC και TAC με ιδιαίτερα καλά αποτελέσματα.

## CLASSY

Το CLASSY [**Conroy06, ConroyACL06, Conroy07**] στηρίχτηκε σε έξι βασικά βήματα για την παραγωγή των περιλήψεων.

***Προετοιμασία εγγράφων - εκκαθάριση προτάσεων:*** Το στάδιο της προετοιμασίας των κειμένων ξεκινά με το διαχωρισμό τους σε προτάσεις (**sentence splitting**). Στη συνέχεια οι προτάσεις κατηγοριοποιούνται, ως υποψήφιες για περίληψη, μη-υποψήφιες αλλά χρήσιμες και μη-υποψήφιες. Ο δεύτερος χαρακτηρισμός, μη-υποψήφιες αλλά χρήσιμες, αφορά εκείνες τις προτάσεις που ενώ δεν θα συμμετάσχουν στη δημιουργία της περίληψης, μπορούν να συνεισφέρουν χρήσιμους όρους-λέξεις στην παραγωγή της. Κατά το επόμενο βήμα, την *εκκαθάριση* (**trimming**) των προτάσεων αφαιρούνται μη σημαντικές πληροφορίες, όπως ημερομηνίες, σχόλια του συγγραφέα αλλά και λέξεις, όπως «also» και «almost» κ.ά., που δεν έχουν μεγάλη πληροφοριακή αξία.

***Επιλογή λέξεων - κλειδιών από την περιγραφή του ζητήματος:*** Κατά την δεύτερη φάση, το σύστημα επιλέγει λέξεις από τη σύντομη *περιγραφή του θέματος* (**topic description**). Αυτή η περιγραφή είναι συνήθως μια πρόταση που αποδίδει το περιεχόμενο των κειμένων από τα οποία θα προκύψει η περίληψη. Οι λέξεις που επιλέγονται περιγράφουν τις *κύριες ιδέες* (**key concepts**) των κειμένων. Συγκεκριμένα, το CLASSY φιλτράρει τις συχνά εμφανιζόμενες λέξεις (**stop words** ) όπως άρθρα («her», «his», «it» κ.ά.), συνδέσμους («and», «also» κ.ά.) και στη συνέχεια, από τις υπόλοιπες λέξεις αφαιρούνται οι καταλήξεις (**stemming**) (π.χ. «fishing», «fisher» και «fished» μετατρέπονται σε «fish»).

***Υπολογισμός σημαντικών όρων:*** Στη συνέχεια, ο αλγόριθμος εντοπίζει τους *σημαντικούς όρους* (**signature terms**) που χρησιμοποιούνται για ένα επόμενο βήμα, αυτό της βαθμολόγησης των προτάσεων του συνόλου των κειμένων απ' τα οποία θα παραχθεί η περίληψη. Οι σημαντικοί όροι είναι αυτοί που εμφανίζονται συχνότερα (με στατιστικά σημαντική διαφορά από άλλους) σε συστάδες κειμένων με κοινό περιεχόμενο. Στον πιο πρόσφατο διαγωνισμό, το CLASSY επέκτεινε το

σύνολο απ' το οποίο ανακτούνταν οι σημαίνοντες όροι. Σε παλαιότερους διαγωνισμούς, το CLASSY χρησιμοποιούσε τη συλλογή κειμένων TREC Novelty, ενώ στον πιο πρόσφατο χρησιμοποιήθηκε μια μεγαλύτερη συλλογή, η AQUAINT, και επιπλέον επετράπη η χρήση λιστών συχνά εμφανιζόμενων λέξεων (stop words) κατά τον υπολογισμό των σημαντικών όρων. Οι τιμές του μέτρου αξιολόγησης ROUGE, που εξηγούνται παρακάτω, αυξήθηκαν όταν χρησιμοποιήθηκε μεγαλύτερη συλλογή κειμένων ως πηγή ανάκτησης σημαντικών όρων.

***Βαθμολόγηση προτάσεων****:* Η βαθμολόγηση των προτάσεων στο CLASSY έγινε χρησιμοποιώντας ένα μέτρο που ονομάστηκε *βαθμός πρόβλεψης (*oracle score*).* Ο βαθμός πρόβλεψης βασίζεται στην κατανομή των *μονογραμμάτων* (unigrams) στα *μοντέλα περίληψης* (model summaries) που συντάχθηκαν από κριτές. Βάσει αυτού οι προτάσεις βαθμολογούνται υπολογίζοντας το κλάσμα των όρων που η κάθε πρόταση έχει κοινούς με υποτιθέμενες (άγνωστες) περιλήψεις ανθρώπων. Αυτό αποτυπώνεται στην παρακάτω παράσταση.

$$w_{qsp}(x) = \frac{1}{|x|}\sum_{t\epsilon T} x(t)\, P_{qsp}\,(t|T), \qquad \text{όπου}$$

Τύπος 1 – Παράσταση oracle score.

Το $w_{qsp}(x)$ είναι ο βαθμός πρόβλεψης της πρότασης *x,* |x| είναι ο αριθμός των μοναδικών όρων της πρότασης *x,* $t$ είναι ένας όρος, $T$ είναι το σύνολο όλων των όρων που χρησιμοποιούνται για το *θέμα* (**topic**) που πραγματεύεται ο αλγόριθμος και *x(t)=1* εάν ο όρος *t* βρίσκεται στην πρόταση *x,* διαφορετικά *x(t)=0.* Ο *βαθμός πρόβλεψης* εκτιμάται και βάσει της πιθανότητας ένας όρος $t$ να συμπεριληφθεί σε ένα *μοντέλο περίληψης* δοθέντος ενός ζητήματος *T,* όπως γίνεται φανερό στην Εικόνα 2. Αυτή η πιθανότητα αποτυπώνεται ως $P_{qsp}\,(t|T)$. Η πιθανότητα αυτή υπολογίζεται χρησιμοποιώντας τις *λέξεις - κλειδιά* του ερωτήματος, τους σημαντικούς όρους και την κατανομή των όρων στη συστάδα των σχετικών κειμένων, ενώ ορίζεται ως :

$$P_{qsp}(t|T) = \frac{1}{4}q_t(T) + \frac{1}{4}s_t(T) + \frac{1}{2}p_t(T), \qquad \text{όπου}$$

Τύπος 2.

$q_t(T)$ = 1 εάν ο όρος $t$ είναι *λέξη - κλειδί του ερωτήματος* για το ζήτημα $T$, διαφορετικά ισούται με το 0, $s_t(T) = 1$ εάν ο όρος $t$ είναι *σημαντικός όρος* για το ζήτημα $T$ διαφορετικά ισούται με το 0 και $p_t(T)$ είναι η πιθανότητα ο όρος $t$ να εμφανίζεται σε μια από τις προτάσεις που πρόκειται να χρησιμοποιηθούν για την περίληψη.

***Αφαίρεση πλεονασμού χρησιμοποιώντας LSI/L1-QR:*** Το προτελευταίο βήμα της διαδικασίας είναι η αποφυγή του *πλεονασμού* στην παραγόμενη περίληψη. Αυτό επιτυγχάνεται χρησιμοποιώντας τον αλγόριθμο Latent Semantic Indexing (LSI) ακολουθούμενο από την L1-norm παραλλαγή του pivoted-QR. Ο pivoted-QR είναι ένας αλγόριθμος αποσύνθεσης ενός πίνακα A (**pivoted-QR decomposition**).

Η παραπάνω διαδικασία περιγράφεται αναλυτικότερα στο [Conroy2006]. Τέλος, αν η τελευταία πρόταση που προστίθεται αυξήσει το μήκος της πρότασης πέρα από το επιθυμητό, τότε αυτή προσαρμόζεται στο επιθυμητό μήκος.

Το CLASSY κατετάγη στο DUC 2007 1ο και 4ο με βάσει τα αυτόματα μέτρα ROUGE-1 ROUGE-2 αντίστοιχα, που αξιολογούν το κατά πόσο το περιεχόμενο των παραγόμενων περιλήψεων είναι όμοιο με εκείνο περιλήψεων που συντάχθηκαν από ανθρώπους.

## Το σύστημα του Οικονομικού Πανεπιστημίου Αθηνών (AUEB2008)

Το 2008, το Οικονομικό Πανεπιστήμιο Αθηνών (Ο.Π.Α.) συμμετείχε [GalMal08] στο Text Analysis Conference (TAC) του 2008 παρουσιάζοντας το δικό του σύστημα για την παραγωγή περιλήψεων από ειδησεογραφικά κείμενα δεδομένης μιας ερώτησης.

Το σύστημα βασίζεται στον αλγόριθμο μηχανικής μάθησης Support Vector Regression (SVR) [ChangLin2001]. Ο αλγόριθμος αυτός εκπαιδεύεται με n διανύσματα εκπαίδευσης, $(x_1, y_1),\ldots,(x_n,y_n)$, όπου το $x_i \in R^n$ και $y_i \in R^n$, και μαθαίνει μια συνάρτηση $f: R^n \to R$ που γενικεύει τα δεδομένα εκπαίδευσης.

Στην περίπτωση του συστήματος του Ο.Π.Α., κάθε ($x_i$, $y_i$) αναπαριστά μια πρόταση. Συγκεκριμένα, τα $x_i$ είναι οι τιμές κάποιων γνωρισμάτων (features) της πρότασης και $y_i$ είναι ένας βαθμός που δείχνει την καταλληλότητά της και το πόσο σχετική είναι με το ερώτημα του χρήστη.

Στα πειράματα του συστήματος για το TAC 2008 χρησιμοποιήθηκαν τα παρακάτω 6 γνωρίσματα:

1. **Sentence position SP(c)**: Το γνώρισμα αυτό εκφράζει τη θέση της πρότασης μέσα στο κείμενο από το οποίο προέρχεται.

   $$SP(c) = \frac{position(c)}{L(c)},\ \text{όπου}$$

   **c** είναι μια πρόταση**, position(c)** είναι η θέση της πρότασης στο έγγραφο (π.χ. 1η, 2η, 3η, ...) και **L(c)** είναι ο συνολικός αριθμός προτάσεων του κειμένου.

2. **Named Entities NE(c).**

   $$NE(c) = \frac{n(c)}{length(c)},\ \text{όπου}$$

   **n(c)** είναι ο αριθμός των ονομάτων οντοτήτων (named entities) στην πρόταση **c.** Τα ονόματα οντοτήτων εντοπίζονται χρησιμοποιώντας την έκδοση 1.1 του σχετικού εργαλείου (named entity recognizer) του πανεπιστημίου του Stanford.[3]

3. **Levenstein distance LD(c,q)**: Είναι η τιμή της απόστασης Levenstein (edit distance) μεταξύ της *ερώτησης* (**query**) **q** και της πρότασης **c**.

4. **Word overlap WO(c,q)**: Εκφράζει την επικάλυψη των λέξεων μεταξύ του ερωτήματος q και της πρότασης c. Οι *συχνά εμφανιζόμενες λέξεις* (stop words) και οι διπλοεμφανιζόμενες λέξεις στις προτάσεις έχουν αφαιρεθεί.

5. **Content word frequency:** Είναι η μέση πιθανότητα εμφάνισης των λέξεων $t_{i../s/}$ μιας πρότασης στη συστάδα των κειμένων (cluster). Η πιθανότητα εμφάνισης μιας λέξης ορίζεται ως:

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

$$p_c(t_i) = \frac{n}{N}, \text{ όπου}$$

n είναι ο αριθμός των εμφανίσεων της λέξης $t_i$ μέσα στη συστάδα και N το σύνολο λέξεων στην συστάδα των κειμένων.

6. **Document frequency:** Είναι η μέση πιθανότητα ένα κείμενο να περιέχει τις λέξεις $t_{i../s/}$ μιας πρότασης. Η πιθανότητα ένα κείμενο να περιέχει μια λέξη ορίζεται ως:

$$p_d(t_i) = \frac{d}{D}, \text{ όπου}$$

d είναι ο αριθμός των κειμένων στα οποία εμφανίζεται η λέξη $t_i$ και D είναι ο συνολικός αριθμός των κειμένων στην συστάδα.

Τα δυο τελευταία γνωρίσματα ορίζονται αναλυτικότερα στο [Schilder2008].

Ως yi χρησιμοποιήθηκε ο μέσος όρος των τιμών δυο μέτρων, του ROUGE-2 και του ROUGE-SU4 [Lin2004] κάθε πρότασης με τις αντίστοιχες *περιλήψεις* (**model summaries**) οι οποίες είχαν δημιουργηθεί από κριτές. Τα ROUGE-2 και ROUGE-SU4 υπολογίζουν ένα βαθμό ομοιότητας δύο κειμένων βάσει των κοινών 2-γραμμάτων λέξεων (word bigrams) [Lin2004].

Το σύστημα εκπαιδεύτηκε χρησιμοποιώντας τις προτάσεις των κειμένων του DUC 2006 καθώς και τις αντίστοιχες περιλήψεις που δημιούργησαν οι κριτές για κάθε σύνολο εγγράφων. Ύστερα, το σύστημα χρησιμοποιήθηκε για να αναθέσει *βαθμούς σχετικότητας* (**relevance scores**) σε κάθε πρόταση των κειμένων του DUC 2007. Για κάθε σύνολο εγγράφων, οι αντίστοιχες προτάσεις ταξινομήθηκαν βάσει των βαθμών σχετικότητάς τους. Ξεκινώντας από τις προτάσεις με την μεγαλύτερη τιμή, το σύστημα πρόσθετε στην περίληψη κάθε πρόταση της οποίας η ομοιότητα με τις προτάσεις που είχαν ήδη περιληφθεί στην περίληψη δεν ξεπερνούσε κάποιο ορισμένο *κατώφλι* (**threshold**). Για τον υπολογισμό της ομοιότητας μεταξύ των προτάσεων χρησιμοποιήθηκε το μέτρο cosine similarity. Επιπλέον, πριν το σύστημα εισαγάγει μια πρόταση στην περίληψη, εφάρμοζε μια διαδικασία απλοποίησης. Κατά την διάρκεια αυτής μια σειρά από φράσεις όπως “in fact”, “however” κ.α διαγράφονταν. Τέλος, οι περιλήψεις περικόπτονταν, ώστε να μην ξεπερνούν σε μέγεθος τις 250 λέξεις, το οποίο είχε τεθεί ως όριο στον διαγωνισμό DUC 2007. Η τιμή του κατωφλιού αποφασίστηκε με πειράματα στα

δεδομένα εκπαίδευσης του DUC 2007. Συγκεκριμένα δοκιμάστηκαν διάφορες τιμές του και κρατήθηκε αυτή η οποία οδήγησε σε περιλήψεις με την μεγαλύτερη ομοιότητα (βάσει των ROUGE-2 και ROUGE-SU4) με τις αντίστοιχες περιλήψεις των κριτών.

Χρησιμοποιώντας το σύστημα εκπαιδευμένο στα δεδομένα του DUC 2006 και ακολουθώντας την ίδια διαδικασία παρήχθησαν περιλήψεις για τα δεδομένα του TAC 2008. Το σύστημα του Ο.Π.Α πέτυχε υψηλές τιμές όσον αφορά τα μέτρα ROUGE-2 και ROUGE-SU4, ενώ κατέλαβε χαμηλότερες θέσεις σε σχέση με τα άλλα συστήματα στην κατάταξη σχετικά με την *γλωσσική ποιότητα* (**linguistic quality**) των περιλήψεων. Αυτό αποδόθηκε στην απουσία αλγορίθμων για την διάταξη και πραγμάτωση (βλ. Ενότητα 2.2) των προτάσεων των περιλήψεων [GalMal08].

### 2.3.2 Παραγωγή περιλήψεων από ιστολόγια.

Στο διαγωνισμό του TAC 2008 εκτός από ειδησεογραφικά κείμενα δόθηκε από τους διοργανωτές και ένα αριθμός συνόλων από ιστολόγια. Για κάθε τέτοιο σύνολο δόθηκαν 1 ή 2 ερωτήματα τα οποία αφορούσαν τις θετικές και αρνητικές γνώμες για κάποιο θέμα (π.χ. «What features do people like about Vista?»). Παρακάτω παρουσιάζεται ένα σύστημα που συμμετείχε στο διαγωνισμό και είχε πολύ καλή επίδοση.

**Το σύστημα του Πολυτεχνείου του Hong Kong**.

Το σύστημα (PolyU) [Li08] επιχειρεί να ανακτήσει τις πιο *αντιπροσωπευτικές προτάσεις* (**salient sentences**) από τα κείμενα. Αυτή η διαδικασία χωρίστηκε σε δυο μέρη. Αρχικά εντοπίστηκαν οι πιο αντιπροσωπευτικές παράγραφοι των κειμένων και στην συνέχεια οι πιο αντιπροσωπευτικές προτάσεις αυτών. Συγκεκριμένα ακολουθήθηκε η παρακάτω διαδικασία:

- **Καθαρισμός των κειμένων από τις HTML ετικέτες:** Το αρχικό *σύνολο κειμένων* (***corpus***) στο TAC 2008 ήταν σε μορφή HTML. Η φάση αυτή

περιελάμβανε την αφαίρεση όλων των ετικετών της HTML, ώστε το κάθε κείμενο να απαλλαχθεί από αυτές και να είναι έτοιμο για επεξεργασία.

- **Διαχωρισμός προτάσεων:** Ύστερα, το κείμενο χωρίστηκε σε προτάσεις με την χρήση του εργαλείου GATE.[4] Οι προτάσεις που προέκυψαν ελέγχθηκαν με την χρήση ενός απλού φίλτρου και όσες ήταν λιγότερες των 6 λέξεων ή είχαν χαμηλή γλωσσική ποιότητα (grammaticality) αφαιρέθηκαν.

- **Βαθμολόγηση προτάσεων:** Σε αυτό το στάδιο, σε κάθε πρόταση που παρέμεινε ύστερα από το προηγούμενο βήμα ανατέθηκε ένας βαθμός. Αυτός ο βαθμός ήταν ο γραμμικός συνδυασμός των τιμών κάποιων γνωρισμάτων. Πιο συγκεκριμένα, τα γνωρίσματα *Centroid* και *SimtoQuery* περιγράφουν την ποιότητα της πληροφορίας που φέρει μια πρόταση σε σχέση με το σύνολο των κειμένων και την ερώτηση του χρήστη, αντίστοιχα [LiDUC07]. Άλλα γνωρίσματα που χρησιμοποιήθηκαν όπως τα *PosSentiment, PosOpinion, NegSentiment, NegOpinion* δηλώνουν την ύπαρξη θετικών ή αρνητικών συναισθημάτων και γνωμών στις προτάσεις. Τέλος, το γνώρισμα *Position* δηλώνει την θέση της πρότασης μέσα στο κείμενο. Αξίζει να σημειωθεί πως πέρα από τα παραπάνω γνωρίσματα, τα οποία είναι στο *επίπεδο πρότασης* (**sentence level**), το PolyU χρησιμοποίησε παρόμοια γνωρίσματα και σε *επίπεδο παραγράφων* (**paragraph level**). Έτσι, τα *ParaCentroid, ParaSimToQuery, ParaOpinion, ParaSentiment* είναι τα αντίστοιχα γνωρίσματα σε επίπεδο παραγράφων και η λειτουργία τους είναι όμοια με αυτή των γνωρισμάτων σε επίπεδο προτάσεων.

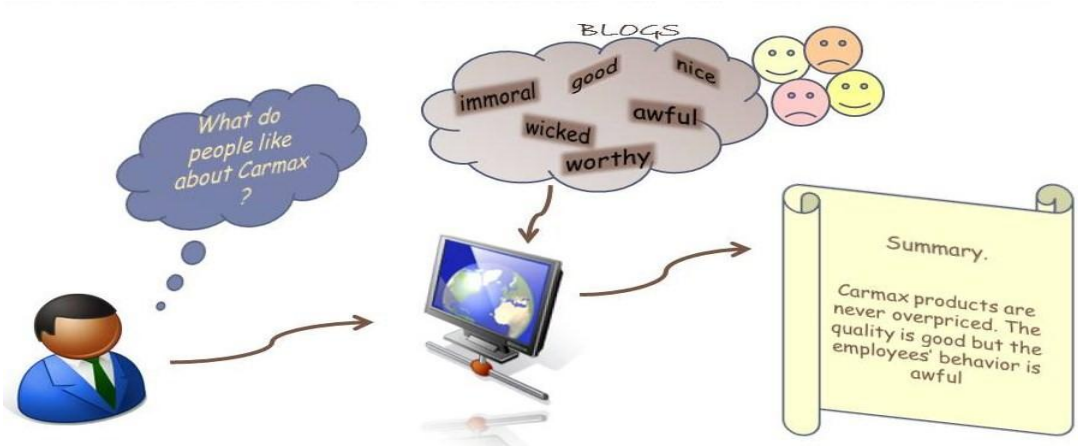
- **Παραγωγή περίληψης:** Για την παραγωγή της περίληψης πρώτα λήφθηκαν υπόψιν τα γνωρίσματα σε επίπεδο παραγράφων, έτσι ώστε να επιλεχθούν εκείνες οι παράγραφοι που περιέχουν τις πιο σχετικές με την ερώτηση του χρήστη απαντήσεις. Στη συνέχεια, οι προτάσεις αυτών των παραγράφων ταξινομήθηκαν βάσει των βαθμών που έλαβαν στην προηγούμενη φάση. Οι βαθμοί όλων των προτάσεων κανονικοποιήθηκαν στο διάστημα 0 έως 1. Τέλος, οι 20 καλύτερες προτάσεις με βαθμούς μεγαλύτερους του 0,3 επιλέχθηκαν για την περίληψη. Για την αποφυγή εμφάνισης πλεονασμού, κάθε πρόταση περιλαμβάνονταν στην περίληψη μόνο όταν δεν έμοιαζε σε μεγάλο

[4] Βλ. http://gate.ac.uk/

βαθμό με προτάσεις που είχαν ήδη επιλεχθεί. Συγκεκριμένα, χρησιμοποιήθηκε ο αλγόριθμος MMR (βλ. ενότητα 2.2).

Το σύστημα κατετάγη πρώτο σε σύνολο 19 συστημάτων που έλαβαν μέρος στο διαγωνισμό σε ό,τι αφορά την ποιότητα των προτάσεων που συμπεριλήφθηκαν στις περιλήψεις [Li2008].

# 3. Το σύστημα και τα πειράματα της εργασίας




Εικόνα 2 – Γενική περιγραφή συστήματος.

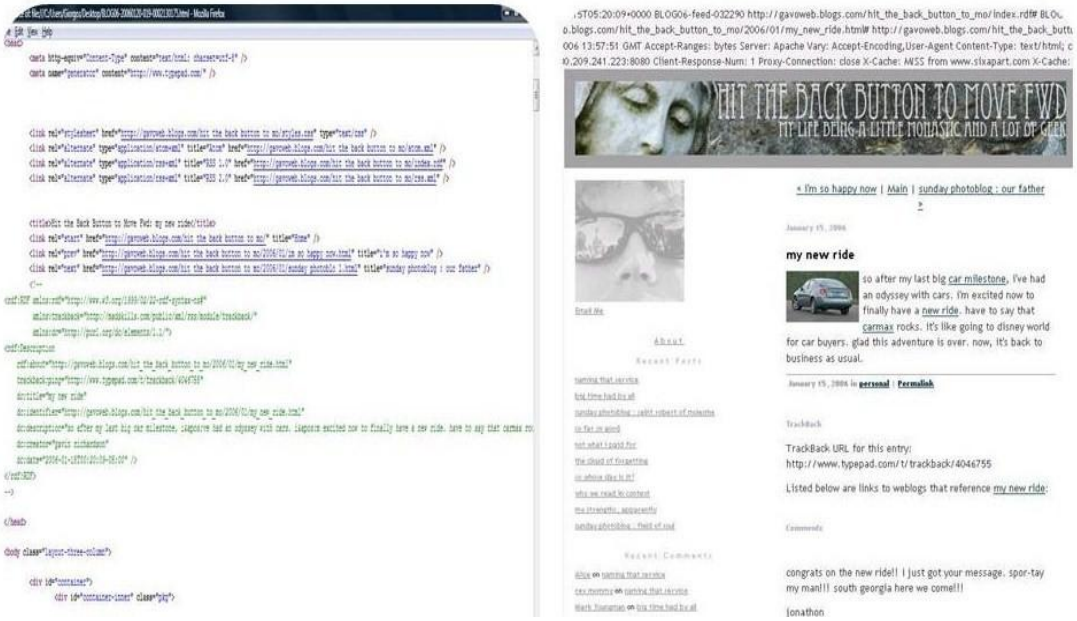
Το σύστημα αυτής της εργασίας είναι επέκταση του συστήματος AUEB08 με το οποίο το Ο.Π.Α. συμμετείχε στο TAC 2008 (βλ. ενότητα 2.3.1). Για την αξιολόγηση του συστήματος εκτελέστηκαν δυο πειράματα. Στο πρώτο πείραμα, ο SVR του AUEB08 εκπαιδεύτηκε σε ειδησεογραφικά κείμενα του διαγωνισμού DUC του 2006 και χρησιμοποιήθηκε για να παράγει περιλήψεις από τα ιστολόγια του διαγωνισμού TAC 2008 (αντί για τα ειδησεογραφικά κείμενα του διαγωνισμού στα οποία είχε αξιολογηθεί το 2008). Στο δεύτερο πείραμα χρησιμοποιήθηκε η τεχνική της διασταυρωμένης επικύρωσης (*cross-validation*) για την εκπαίδευση και στην συνέχεια για την παραγωγή περιλήψεων από τα ιστολόγια του TAC 2008. Τα αποτελέσματα που προέκυψαν συγκρίθηκαν με άλλα συστήματα που συμμετείχαν στην παραγωγή περιλήψεων από ιστολόγια (**opinion summarization track**) του διαγωνισμού TAC 2008.

Εικόνα 3 - Δομή Συστήματος AUEBop10

## 3.1 Προεπεξεργασία ιστολογίων.

Πριν χρησιμοποιηθούν, τα ιστολόγια υπέστησαν προεπεξεργασία με σκοπό την αφαίρεση HTML ετικετών, καταλόγων επιλογών κ.λ.π. που δεν έχουν αξία για τους σκοπούς της παραγωγής περιλήψεων, ώστε να ανακτηθεί μόνο το «πραγματικό» κείμενο. Η προεπεξεργασία των ιστολογίων έγινε σε δύο στάδια (βλ. Εικόνα 3):

**Στάδιο 1ο**: Χρησιμοποιήθηκε το εργαλείο ανοιχτού λογισμικού NReadability, για να καθαρίσει τα ιστολόγια από συνδέσμους, πλαϊνές στήλες (sidebar), κώδικα javascript, κ.λ.π..[5]



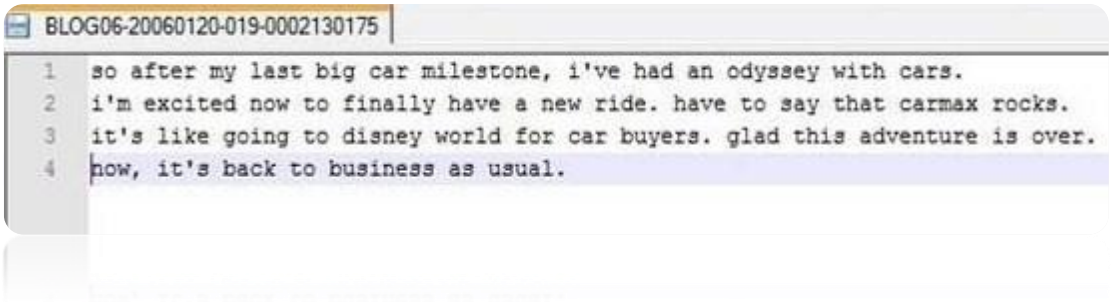

**Εικόνα 4 - Αριστερά ο πηγαίος κώδικας ενός από τα ιστολόγια (raw blog). Δεξιά η πραγματική εικόνα του ιστολογίου στο πρόγραμμα περιήγησης (browser).**

**Στάδιο 2ο :** Τα αρχεία που προέκυψαν από το παραπάνω στάδιο περιείχαν ακόμα κώδικα HTML, όπως <p>, <html>, <body>,<img> αλλά μόνο για λόγους μορφοποίησης του κειμένου (βλ. Εικόνα 5).

[5] Βλ. http://code.google.com/p/nreadability/

Εικόνα 5 - Αριστερά το αποτέλεσμα της επεξεργασίας του ιστολογίου από το NReadability. Δεξιά το αποτέλεσμα στον περιηγητή.

Για τον καθαρισμό των ιστολογίων από αυτά αναπτύχθηκε λογισμικό, που βασίστηκεστη βιβλιοθήκη ανοιχτού λογισμικού Jericho.[6] Το αποτέλεσμα ήταν να ανακτηθεί στην πλειονότητα των περιπτώσεων εντελώς *καθαρό κείμενο*, όπως φαίνεται στην παρακάτω εικόνα.




Εικόνα 6 - Το κείμενο που ανακτήθηκε μετά την επεξεργασία (clean blog).

## 3.2 Πειράματα

Η αξιολόγηση των περιλήψεων που παράχθηκαν έγινε με την ίδια διαδικασία όπως στο διαγωνισμό TAC 2008. Η διαδικασία αυτή ξεκινά με τη χειρωνακτική αντιστοίχιση των προτάσεων της κάθε περίληψης στις *ιδανικές προτάσεις* (**nuggets**). Οι ιδανικές προτάσεις είναι τμήματα του κειμένου των αρχικών ιστολογίων ή παραφράσεις τους οι οποίες απαντούν στην ερώτηση που τίθεται

[6] http://jericho.htmlparser.net/docs/index.html

από το χρήστη και οι οποίες έχουν εντοπιστεί από ανθρώπους-κριτές. Κάθε πρόταση της περίληψης εξετάζεται για να διαπιστωθεί αν φέρει τις ίδιες πληροφορίες με κάποια από τις δοσμένες ιδανικές προτάσεις. Κάθε ιδανική πρόταση φέρει έναν προκαθορισμένο *βάρος* (**nugget score**) στην κλίμακα [0, 1] που δείχνει πόσο σημαντική είναι.

Όλα τα συστήματα συγκρίθηκαν χρησιμοποιώντας το μέτρο F-score. Αυτό συνδυάζει άλλα δύο μέτρα, την *ανάκληση* (**recall**) και την *ακρίβεια* (**precision**).

$$F - score = \frac{(\beta^2 + 1) * R * P}{\beta^2 * P + R}$$

Με R συμβολίζεται η ανάκληση, με P η ακρίβεια και β είναι ένας παράγοντας που καθορίζει το βάρος των P και R στον υπολογισμό του F-score. Χρησιμοποιούμε την τιμή 1 για το β (ίσο βάρος στα P και R), καθώς η ίδια τιμή χρησιμοποιήθηκε και κατά την αξιολόγηση των συστημάτων του TAC 2008.

Ο τύπος για τον υπολογισμό της ανάκλησης είναι ο παρακάτω:

$$R = \frac{r}{\boldsymbol{R}}$$

Με r συμβολίζεται το άθροισμα των προκαθορισμένων βαρών των ιδανικών προτάσεων που αντιστοιχούν σε κάποια πρόταση της υπό αξιολόγηση περίληψης, ενώ με **R** συμβολίζεται το άθροισμα των προκαθορισμένων βαρών όλων των ιδανικών προτάσεων.

Ο τύπος με τον οποίο υπολογίζεται η ακρίβεια είναι ο παρακάτω:

$$P = 1 \text{ , } \alpha\nu \; l < A$$

$$P = 1 - \frac{l-A}{l}, \text{ αλλιώς.}$$

Το $l$ είναι το πλήθος των χαρακτήρων της περίληψης αν εξαιρεθούν οι χαρακτήρες των διαστημάτων μεταξύ των λέξεων και της αλλαγής γραμμής. Το A είναι ένα *περιθώριο* (**allowance**) και ορίζεται ως εξής :

$$A = C * a$$

C είναι ο αριθμός χαρακτήρων των ιδανικών προτάσεων (C=100), ενώ α είναι ο αριθμός των ιδανικών προτάσεων που ταιριάζουν με κάποια από τις προτάσεις της περίληψης. Ουσιαστικά η ακρίβεια επιχειρεί να αξιολογήσει το κατά πόσον η

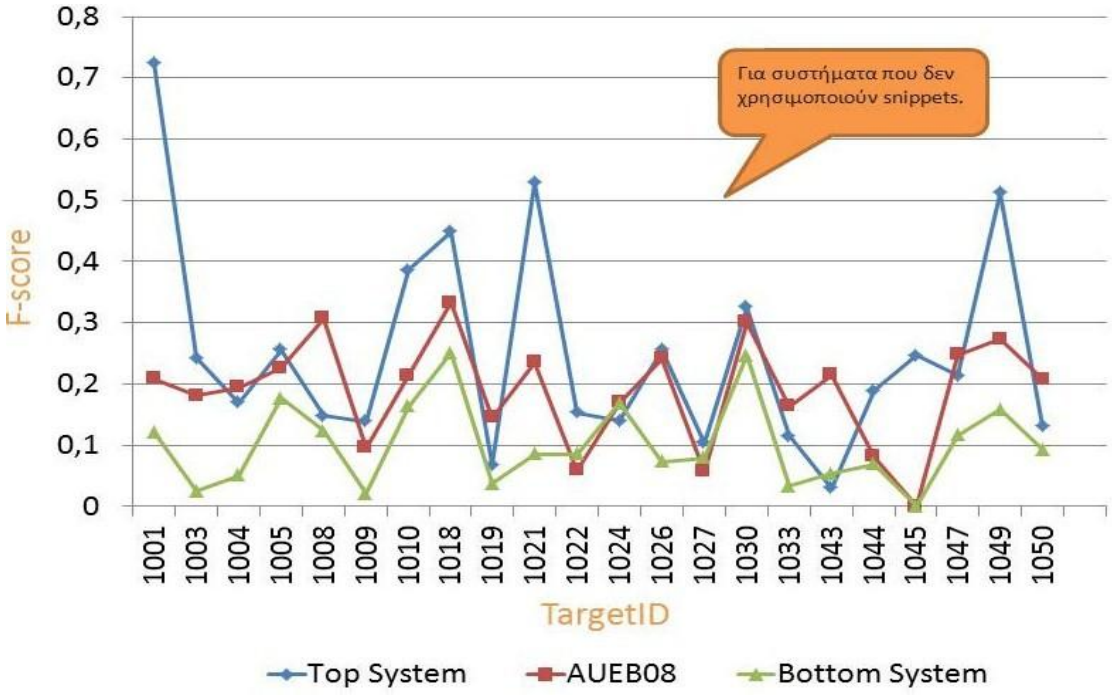
περίληψη περιλαμβάνει περιττές πληροφορίες, μετρώντας το κατά πόσον υπερβαίνει το χώρο των ιδανικών προτάσεων στις οποίες αντιστοιχεί.

Πριν την παρουσίαση των πειραμάτων πρέπει να διευκρινιστεί πως σε κανένα από τα δύο πειράματα δεν χρησιμοποιήθηκαν οι *προτάσεις-κλειδιά* (**snippets**) για την εκπαίδευση του συστήματος. Οι προτάσεις-κλειδιά είναι προτάσεις που έχουν εξαχθεί από τα ιστολόγια και οι οποίες αποτελούν απαντήσεις στην ερώτηση κάθε θεματικής ενότητας. Διατέθηκαν στους διαγωνιζόμενουςτου TAC 2008 από τους διοργανωτές του διαγωνισμού, αλλά στην πράξη δεν θα ήταν διαθέσιμες, παρά μόνο αν ένα σύστημα παραγωγής περιλήψεων χρησιμοποιείτο σε συνδυασμό με ένα (ιδανικό) σύστημα ερωταποκρίσεων (**question-answering system**).

## Πείραμα 1.

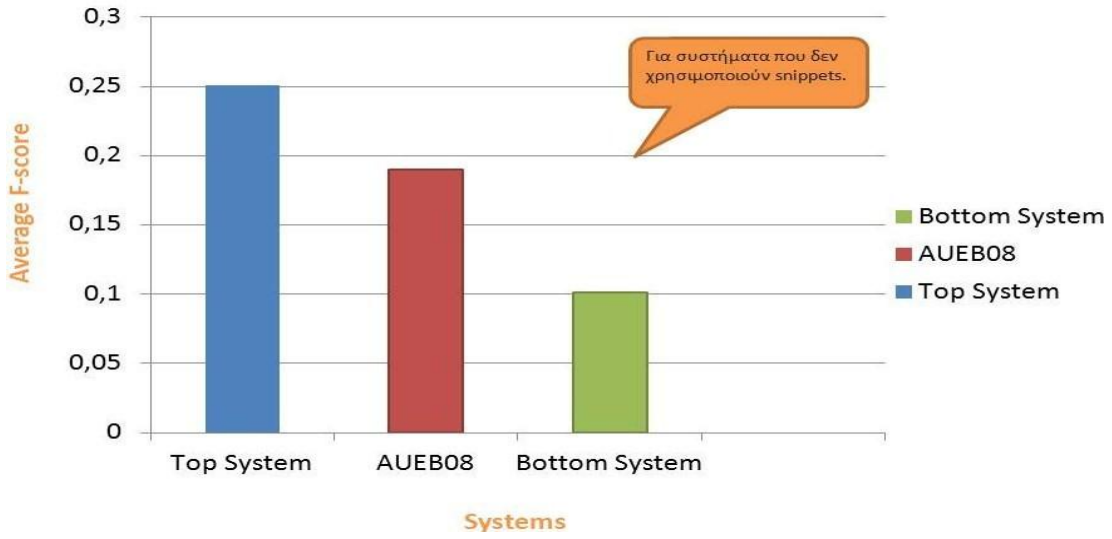
Στα πλαίσια του πρώτου πειράματος, το σύστημα AUEB08 που παρουσιάστηκε στην ενότητα 2.3.1 εκπαιδεύτηκε στα δεδομένα του DUC 2006 και αξιολογήθηκε στα ιστολόγια του TAC 2008. Το σύνολο των δεδομένων του DUC 2006 περιελάμβανε ειδησεογραφικά κείμενα χωρισμένα σε 45 *θεματικές ενότητες (45 σύνολα κειμένων)* πάνω στα οποία το σύστημα εκπαιδεύτηκε λαμβάνοντας υπόψιν τις αντίστοιχες *ερωτήσεις* (**queries**) καθώς και τις περιλήψεις που είχαν δημιουργηθεί από κριτές (model summaries). Τα *δεδομένα ελέγχου* (testing), αποτελούνταν από ιστολόγια χωρισμένα σε 22 θεματικές ενότητες. Τα αποτελέσματα του πειράματος παρουσιάζονται στα παρακάτω διαγράμματα.

Η εικόνα 9 απεικονίζει το F-score που αντιστοιχεί στις περιλήψεις των 22 θεματικών ενοτήτων για το καλύτερο σύστημα του TAC 2008, το σύστημα AUEB08 (περιλήψεις 850 λέξεων) και το τελευταίο στην κατάταξη σύστημα.




Εικόνα 7 - Το F-score για κάθε topic των blogs για τρία συστήματα.

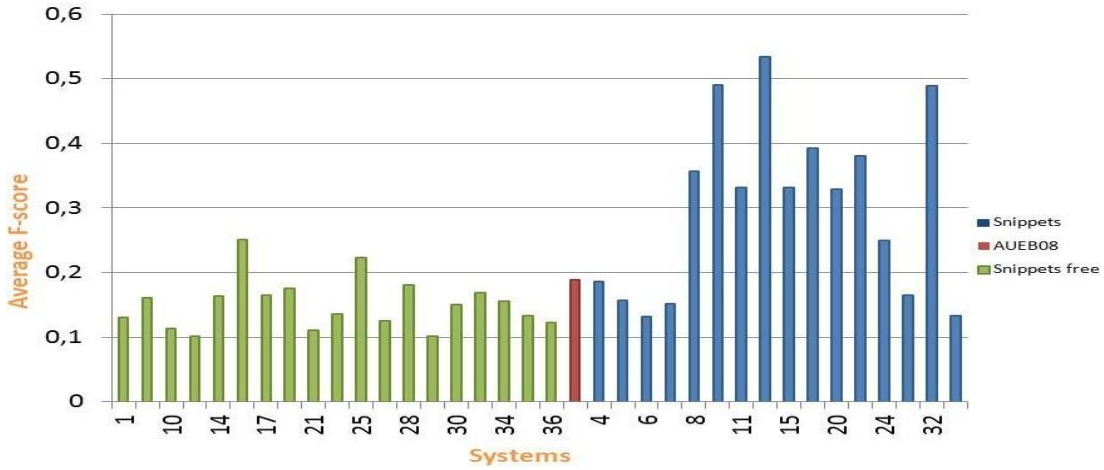
Ο μέσος όρος (για όλες τις θεματικές ενότητες) των F-score των παραπάνω τριών συστημάτων φαίνεται στην εικόνα 8.




Εικόνα 8 - Η κατάταξη του AUEB08 με Average F-score 0,19

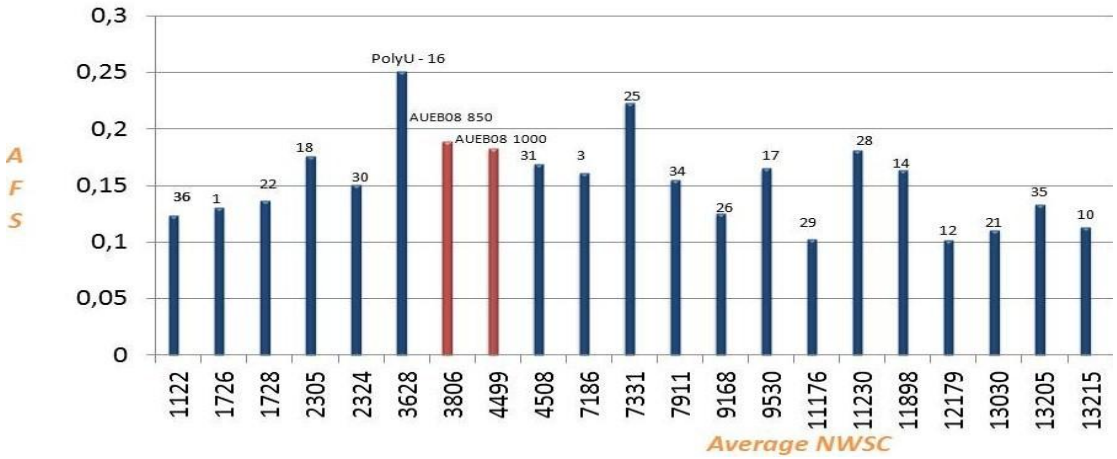
Η κατάταξη του AUEB08 μεταξύ όλων των συστημάτων που συμμετείχαν στο TAC 2008 φαίνεται στο παρακάτω διάγραμμα. Στο παρόν πείραμα, το AUEB08 παρήγαγε περιλήψεις μεγέθους 850 λέξεων. Πολλά από τα συστήματα του διαγωνισμού χρησιμοποίησαν ως επιπλέον πληροφορία τις προτάσεις-κλειδιά

(snippets). Όπως φαίνεται και στο διάγραμμα, αυτά τα συστήματα είχαν πολύ καλύτερες επιδόσεις σε σχέση με τα υπόλοιπα.



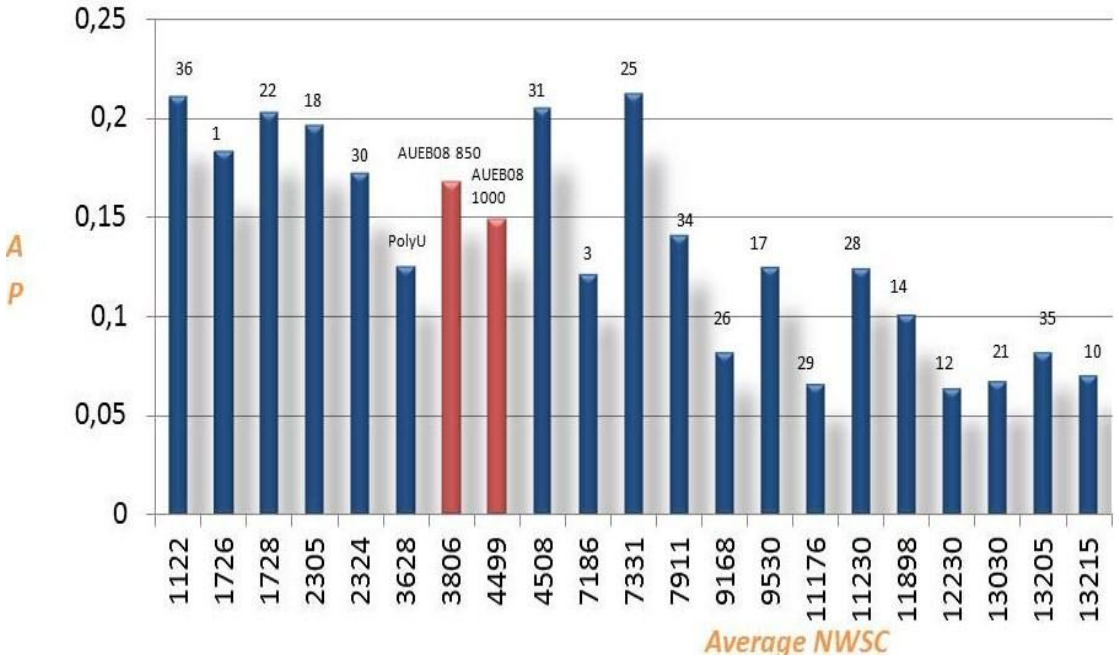

Εικόνα 9 - Η κατάταξη του AUEB08 μεταξύ όλων των συστημάτων του διαγωνισμού.

Το F-score (**average F-score**) του κάθε συστήματος, επηρεάζεται σημαντικά από το μέγεθος (σε χαρακτήρες) της κάθε περίληψης που δημιούργησε. Παρακάτω παρουσιάζονται τρία διαγράμματα που απεικονίζουν τη σχέση του μέσου αριθμού χαρακτήρων των περιλήψεων κάθε συστήματος με κάθε ένα από τα μέτρα *F-score*, *ακρίβεια* και *ανάκληση*. Τα παρακάτω συστήματα, 20 στο σύνολό τους, δεν χρησιμοποίησαν τις *προτάσεις-κλειδιά* (**snippets**).

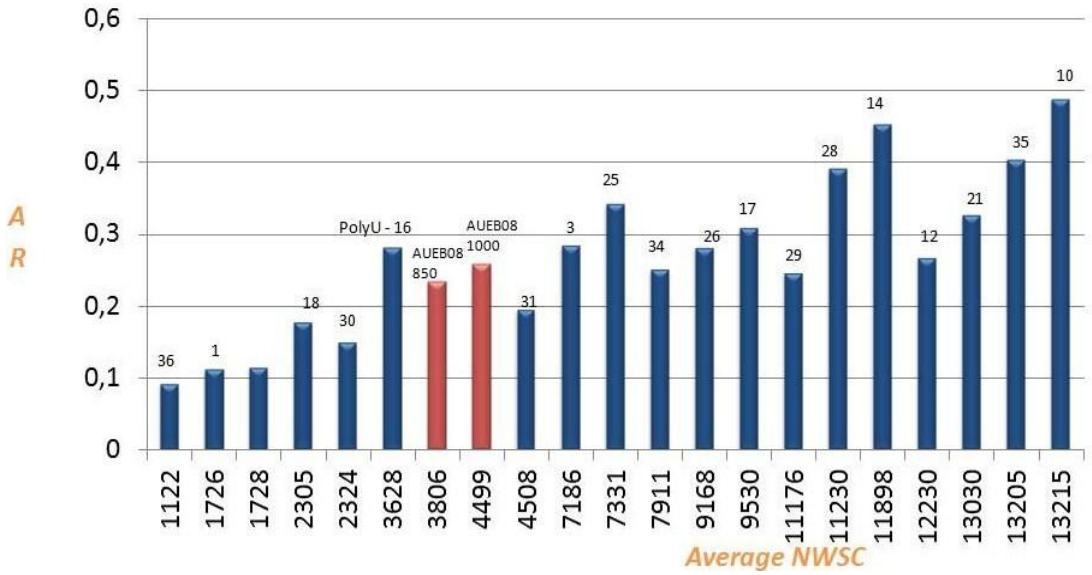



Εικόνα 10 - Το μέσο F-score σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.

Στην παραπάνω εικόνα διακρίνεται το μέσο F-score του πρώτου στην κατάταξη συστήματος (**PolyU**), καθώς και το μέσο F-score του AUEB08 για περιλήψεις των 850 και 1000 λέξεων. Στα παρακάτω διαγράμματα φαίνεται πως αυξάνοντας το επιτρεπόμενο μήκος των περιλήψεων από 850 σε 1000 λέξεις, αυξάνεται η ανάκληση και μειώνεται η ακρίβεια, όπως είναι αναμενόμενο.




Εικόνα 11 – Η μέση ακρίβεια σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.




Εικόνα 12 - Η μέση ανάκληση σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.

## Πείραμα 2.

Στα πλαίσια του δεύτερου πειράματος ενσωματώθηκαν επιπλέον γνωρίσματα στα διανύσματα εκπαίδευσης με σκοπό την ανάκτηση από τα ιστολόγια προτάσεων που εκφράζουν θετικές ή αρνητικές απόψεις και οι οποίες δεν ανακτήθηκαν από

το σύστημα του προηγούμενου πειράματος. Πιο συγκεκριμένα τα γνωρίσματα που προστέθηκαν είναι:

- **Absolute Sentiment Score:** Το γνώρισμα αυτό δηλώνει το μέσο απόλυτο βαθμό *συναισθηματικότητας* (**sentiment**) των λέξεων μιας πρότασης. Ο βαθμός αυτός προκύπτει από το άθροισμα των βαθμών συναισθηματικότητας των λέξεων διαιρούμενο προς το πλήθος αυτών των λέξεων. Ο βαθμός συναισθηματικότητας είναι μια τιμή στο διάστημα [-1,1] και δηλώνει κατά πόσο μια λέξη εκφράζει θετική, αρνητική ή ουδέτερη γνώμη. Κατά τον υπολογισμό εξαιρούνται οι προτάσεις που έχουνε μηδενικό βαθμό *συναισθηματικότητας.* Παρακάτω δίνεται ένα παράδειγμα υπολογισμού αυτού του βαθμού.

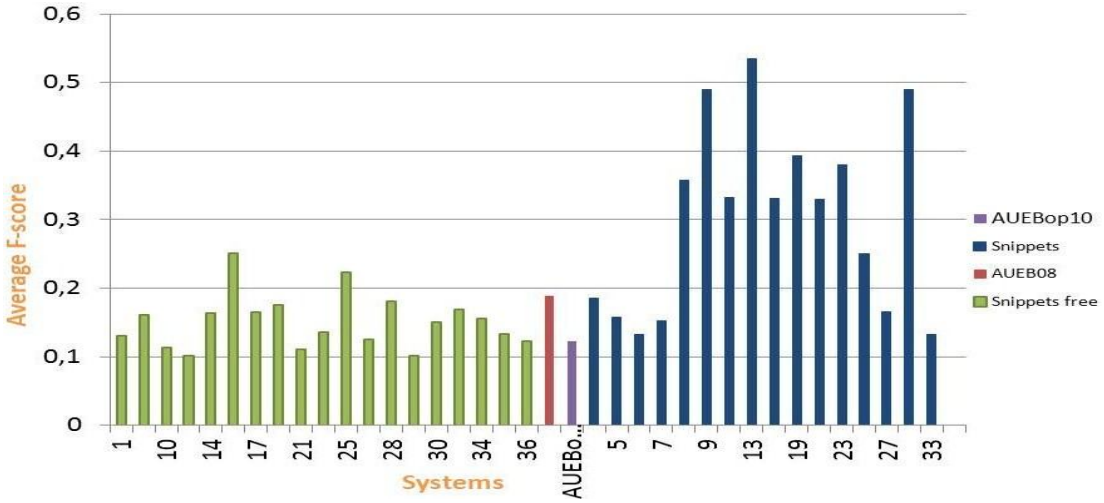
Εικόνα 13 - Παράδειγμα ουδέτερων λέξεων και λέξεων με συναισθηματικότητα.

- **Sentiment Score Percentage:** Είναι το ποσοστό των λέξεων της πρότασης που έχουν είτε αρνητικό είτε θετικό βαθμό συναισθηματικότητας,.
- **Maximum Sentiment Score:** Δηλώνει το μέγιστο βαθμό συναισθηματικότητας των λέξεων της πρότασης.
- **Minimum Sentiment Score:** Δηλώνει τον ελάχιστο βαθμό συναισθηματικότητας των λέξεων της πρότασης.

Ο βαθμός συναισθηματικότητας των λέξεων ανακτήθηκε από το SentiWordNet [SentiWN], ένα λεξικό το οποίο αντιστοιχεί βαθμούς συναισθηματικότητας σε λέξεις.

Λόγω της έλλειψης ιστολογίων και αντίστοιχων περιλήψεων που θα μπορούσαν να χρησιμοποιηθούν για την εκπαίδευση του συστήματος, επιλέχθηκε η μέθοδος της *διασταυρωμένης επικύρωσης* (**cross validation**). Το διαθέσιμο σύνολο των

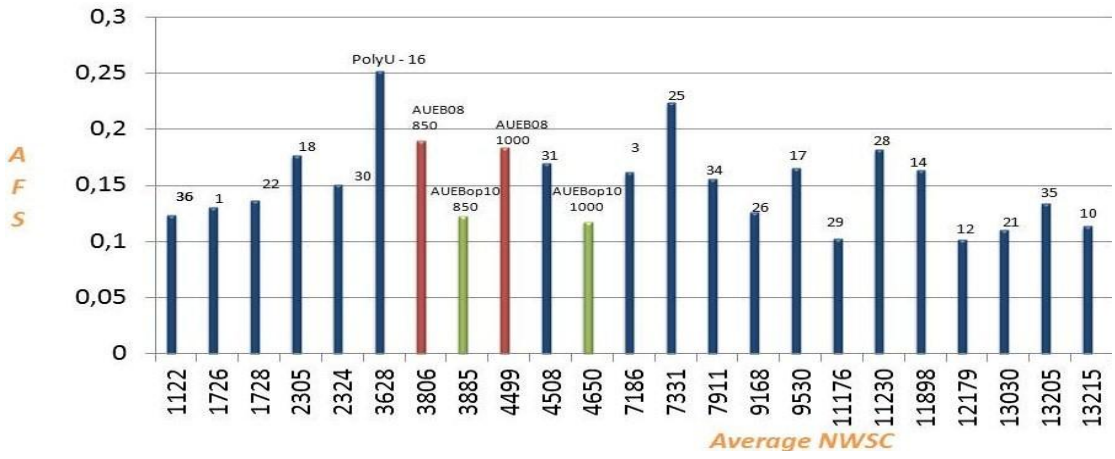
ιστολογίων του TAC 2008 χωρίστηκε σε 10 μέρη. Σε κάθε επανάληψη της διαδικασίας, το σύστημα εκπαιδευόταν στα 9/10 των ιστολογίων και ο *έλεγχος* (**testing**) γινόταν στο 1/10 που είχε εξαιρεθεί από την εκπαίδευση. Κατά την εκπαίδευση, ως μοντέλο περίληψης του κάθε συνόλου ιστολογίων χρησιμοποιήθηκε το αντίστοιχο σύνολο ιδανικών προτάσεων (**nuggets**). Τα αποτελέσματα παρουσιάζονται στα επόμενα διαγράμματα. Το AUEBop10 είναι το AUEB08 με τα πρόσθετα γνωρίσματα.




Εικόνα 14 - Σύγκριση AUEBop10 με τα υπόλοιπα συστήματα του διαγωνισμού.

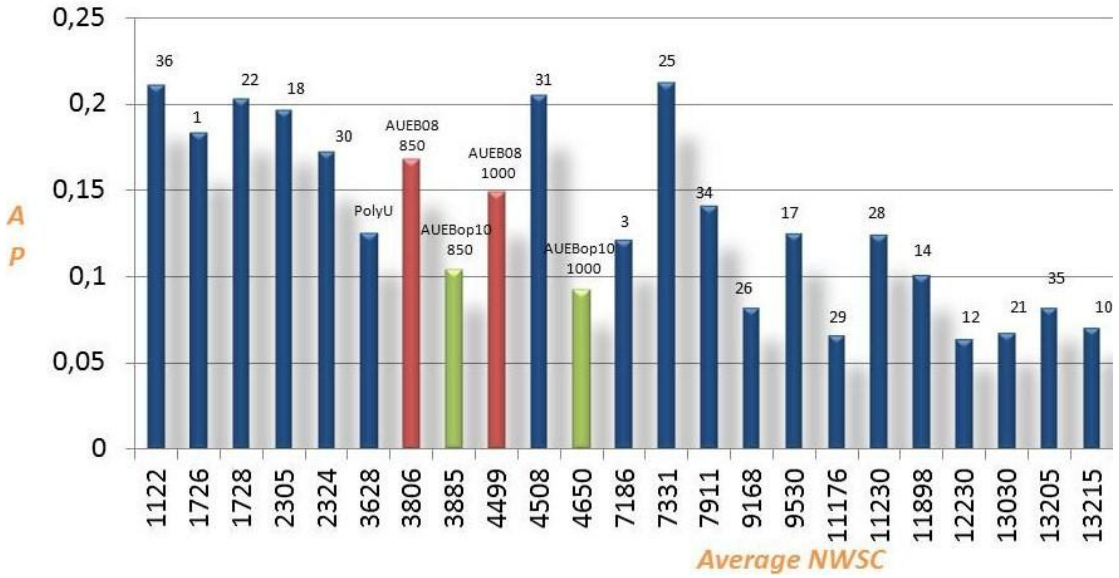
Το F-score που εμφανίζεται στην παραπάνω εικόνα για τα συστήματα AUEB08 και AUEBop10, αφορά προτάσεις μεγέθους 850 λέξεων.

Παρακάτω παρουσιάζονται τρία διαγράμματα που απεικονίζουν την σχέση του μέσου αριθμού χαρακτήρων των περιλήψεων κάθε συστήματος με κάθε ένα από τα μέτρα *F-score*, *ακρίβεια* και *ανάκληση*.

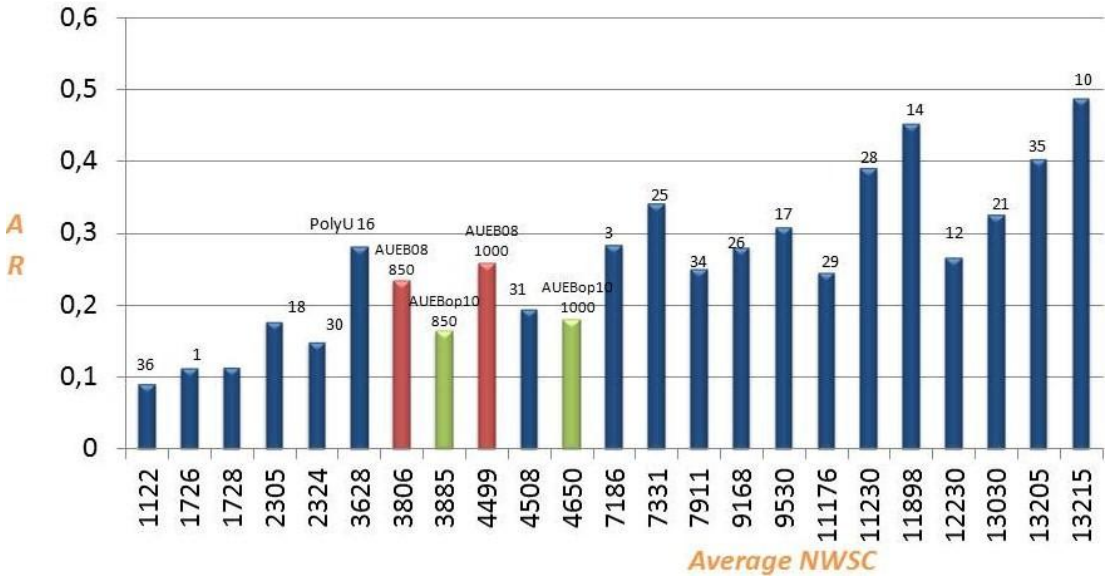



Εικόνα 15- Το average F-score σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.

Στο παραπάνω διάγραμμα παρουσιάζεται το μέσο F-score όλων των συστημάτων, μεταξύ των οποίων είναι το AUEB08 και AUEBop10 για μεγέθη περιλήψεων 850 και 1000 λέξεων.




Εικόνα 16 - Η μέση ακρίβεια σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.




Εικόνα 17 - Η μέση ανάκληση σε σχέση με το μέσο μέγεθος των περιλήψεων κάθε συστήματος.

## Παρατηρήσεις

Βασιζόμενοι στο πρώτο πείραμα, διαπιστώνουμε πως το AUEB08 πετυχαίνει υψηλό F-score και κατατάσσεται 3ο, για μέγεθος περίληψης 850 λέξεων, μεταξύ

των διαγωνιζόμενων συστημάτων στα ιστολόγια. Αν και τα δεδομένα στα οποία εκπαιδεύτηκε είναι ειδησεογραφικά κείμενα, φαίνεται πως τα γνωρίσματα που χρησιμοποιεί είναι ιδιαίτερα αποτελεσματικά, ώστε να μπορεί να παράγει πολύ καλές περιλήψεις και από ιστολόγια.

Όσον αφορά το δεύτερο πείραμα, παρότι προστέθηκαν νέα γνωρίσματα στο AUEB08, αυτό δεν κατάφερε να βελτιώσει τα αποτελέσματά του. Η προσθήκη νέων γνωρισμάτων είχε ως στόχο το νέο σύστημα (AUEBop10) να μπορεί να διακρίνει θετικές ή αρνητικές γνώμες στα ιστολόγια. Ωστόσο, τα αποτελέσματα ήταν χειρότερα, συγκρινόμενα με εκείνα του AUEB08. Ένας πιθανός λόγος για την αδυναμία του AUEBop10 να βελτιώσει τα αποτελέσματα είναι ίσως ότι δεν είχε στην διάθεσή του αρκετά διανύσματα εκπαίδευσης. Τα διανύσματα εκπαίδευσης που παρήχθησαν από τα δεδομένα του DUC 2006 (και στα οποία εκπαιδεύτηκε το AUEB2008) είναι σχεδόν διπλάσια από τα διανύσματα εκπαίδευσης (σε κάθε επανάληψη της διασταυρωμένης επικύρωσης) που χρησιμοποιήθηκαν για την εκπαίδευση του AUEBop10. Ένας άλλος πιθανός λόγος είναι η ύπαρξη μόνο ενός μοντέλου περίληψης για κάθε θεματική ενότητα στα δεδομένα των ιστολογίων. Αντίθετα, στα δεδομένα του DUC 2006 (όπου εκπαιδεύτηκε το AUEB2008) υπήρχαν τέσσερα μοντέλα περίληψης ανά θεματική ενότητα.

# 4. Συμπεράσματα και μελλοντικές κατευθύνσεις

Στη διάρκεια αυτής της εργασίας επεκτάθηκε ένα σύστημα παραγωγής εστιασμένων περιλήψεων από ειδησεογραφικά κείμενα, ώστε να είναι δυνατόν να χρησιμοποιηθεί για την παραγωγή περιλήψεων από ιστολόγια. Το νέο σύστημα κατατάσσεται 3ο ως προς την ποιότητα των παραγόμενων περιλήψεων από ιστολόγια, μεταξύ των συστημάτων που συμμετείχαν στο σχετικό διαγωνισμό του συνεδρίου TAC-2008, παρ' όλο που εκπαιδεύτηκε σε ειδησεογραφικά κείμενα αντί για ιστολόγια.

Επιχειρήθηκε, επίσης, η περαιτέρω επέκταση του συστήματος, μέσω της προσθήκης νέων γνωρισμάτων που στόχο είχαν τον εντοπισμό προτάσεων που εκφράζουν θετικές ή αρνητικές γνώμες. Τα πειραματικά αποτελέσματα, όμως, με τα πρόσθετα γνωρίσματα ήταν χειρότερα, συγκρινόμενα με τα αποτελέσματα του συστήματος χωρίς τα πρόσθετα γνωρίσματα. Τα χειρότερα αποτελέσματα της δεύτερης επέκτασης ίσως οφείλονται στο ότι όταν χρησιμοποιήθηκαν τα πρόσθετα γνωρίσματα το σύστημα εκπαιδεύτηκε σε ιστολόγια που παρείχαν λιγότερα δεδομένα εκπαίδευσης, συγκρινόμενα με τα ειδησεογραφικά κείμενα στα οποία εκπαιδεύτηκε το σύστημα χωρίς τα πρόσθετα γνωρίσματα. Επίσης, τα ειδησεογραφικά κείμενα εκπαίδευσης παρείχαν τέσσερις χειρωνακτικά κατασκευασμένες περιλήψεις ανά θέμα, έναντι μίας των δεδομένων των ιστολογίων.

Η προσθήκη περισσοτέρων περιλήψεων στα κείμενα εκπαίδευσης ιστολογίων, η προσθήκη κειμένων εκπαίδευσης από περισσότερα ιστολόγια και η ενσωμάτωση πιο περίπλοκων μεθόδων εντοπισμού γνωμών θα μπορούσαν ενδεχομένως να βελτιώσουν τις επιδόσεις του συστήματος, όταν χρησιμοποιούνται τα πρόσθετα γνωρίσματα. Μια εναλλακτική προσέγγιση θα ήταν να χρησιμοποιηθούν δύο διαφορετικά στάδια επεξεργασίας. Το πρώτο θα απέρριπτε, χρησιμοποιώντας μόνο τα πρόσθετα γνωρίσματα, προτάσεις που δεν εκφράζουν κάποια γνώμη ή συναίσθημα. Το δεύτερο στάδιο θα χρησιμοποιούσε το υπάρχον σύστημα εκπαιδευμένο σε ειδησεογραφικά κείμενα, χωρίς τα πρόσθετα γνωρίσματα, προκειμένου να επιλέξει τις καλύτερες μεταξύ των προτάσεων που θα είχαν απομείνει μετά το πρώτο στάδιο.

# Βιβλιογραφία

**[SentiWN]** A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", Proceedings of LREC, 2006.

**[ChangLin2001]** C.C Chang and C.J. Lin. "LIBSVM: A library for Support Vector Machines", 2001.

**[Conroy06]** J. M. Conroy D. P. O'Leary, J. D. Schlesinger. "CLASSY Arabic and English multi-document summarization". Multi-Lingual Summarization Evaluation 2006, 2006.

[**ConroyACL06**] J. M. Conroy D J. D. Schlesinger. P. O'Leary, Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. ACL, 2006.

**[Conroy07]** J. M. Conroy J. D. Schlesinger, D. P. O'Leary. "CLASSY 2007 at DUC 2007". Document Understanding Conference, 2007.

**[GalMal08]** D. Galanis and P. Malakasiotis, "AUEB at TAC 2008". Text Analysis Conference, 2008.

**[JurMartin08]** D. Jurafsky and J. Martin. "Speech and Language Processing", 2nd. edition. Pearson, 2008.

**[Li08]** W. Li, Y. Ouyang, Y. Hu, F. Wei. "PolyU at TAC 2008". Text Analysis Conference 2008.

**[LiDUC07]**. S. Li, Y. Ouyang, W. Wang, B. Sun. Multi-document Summarization Using Support Vector Regression. Document Understanding Conference, 2007.

**[Lin2004]** C.W. Lin. "ROUGE: A package for automatic evaluation of summaries". ACL-04 Workshop: Text Summarization Branches Out, 2004.

**[Schilder2008]** F. Schilder and K. Ravikumar. "FastSum: Fast and Accurate Query-based Multi-document Summarization". ACL, 2008.