ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ ΚΑΙ ΓΛΩΣΣΙΚΑ ΕΡΓΑΛΕΙΑ

Opinion Mining

Opinion Mining

- Συνώνυμο: Sentiment Analysis
- Ορισμός:
  - Ανάλυση κειμένων που αναφέρονται σε μια οντότητα/αντικείμενο
  - Εντοπισμός συναισθημάτων ή απόψεων για το αντικείμενο που εκφράζονται γραπτώς
  - Εξαγωγή συμπεράσματος για το αν είναι αρνητικά, θετικά ή ουδέτερα.
- Στη διαδικασία εμπλέκεται μια πλειάδα NLP τεχνικών

Opinion Mining - Example

- Έστω το κείμενο:
  - (1) I bought a phone a few days ago.
  - (2) It was such a nice phone.
  - (3) The touch screen was really cool.
  - (4) The voice quality was clear too.
  - (5) Although the battery life was not long, that is ok for me.
  - (6) However, my mother was mad with me as I did not tell her before I bought it.
  - (7)She also thought the phone was too expensive, and wanted me to return it to the shop.
- Τι έχουμε στόχο να εξάγουμε από τον παραπάνω σχολιασμό?

Opinion Mining - Example

- Θετικά/Αρνητικά/Ουδέτερα συναισθήματα:
  - (1) I bought a phone a few days ago.
  - (2) It was such a nice phone.
    - Άποψη του συγγραφέα για το τηλέφωνο.
  - (3) The touch screen was really cool.
    - Άποψη του συγγραφέα για την οθόνη αφής
  - (4) The voice quality was clear too.
    - Άποψη του συγγραφέα για την ποιότητα ήχου
  - (5) Although the battery life was not long, that is ok for me.
    - Άποψη του συγγραφέα για τη διάρκεια της μπαταρίας
  - (6) However, my mother was mad with me as I did not tell her before I bought it.
    - Άποψη της μητέρας για τον συγγραφέα
  - (7)She also thought the phone was too expensive, and wanted me to return it to the shop.
    - Άποψη της μητέρας για το τηλέφωνο
- Για κάθε άποψη μας ενδιαφέρει:
  - Σε ποιόν ανήκει
  - Για ποιο πράγμα εκφράζεται
  - Η πολικότητά της (αρνητική/θετική/ουδέτερη)

Βασικές Έννοιες - Αντικείμενο

- Ένα αντικείμενο **ο** είναι μια οντότητα που μπορεί να αντιπροσωπεύει προϊόν, πρόσωπο, γεγονός, οργανισμό ή θέμα.
- Συνδέεται με ένα ζεύγος (Τ,Α) όπου
  - Τ είναι μια ιεραρχία(δέντρο) συστατικών ή μερών
    - Δέντρο γιατί τα συστατικά ενός συστατικού ανήκουν επίσης στο αντικείμενο
  - Α είναι ένα σύνολο γνωρισμάτων.
- Στο παράδειγμα:
  - Αντικείμενο **ο**: το κινητό
  - Τ = {οθόνη αφής, μπαταρία, ...}
  - Α= {ποιότητα ήχου, διάρκεια μπαταρίας, κόστος, ...}

Βασικές Έννοιες - Features

- Μπορεί να εκφραστεί άποψη για:
  - Το αντικείμενο
    - "It was a nice phone."
  - Ένα συστατικό του
    - "The touch screen was cool."
  - Γνωρίσματα του αντικειμένου
    - "The voice quality was good."
  - Γνωρίσματα των συστατικών
    - "The battery life was not long."
- Στην πράξη χρησιμοποιούμε τον όρο *features* για να εκφράσουμε το σύνολο των συστατικών και των γνωρισμάτων.
  - Στα features συμπεριλαμβάνεται και το ίδιο το αντικείμενο.
  - Κάθε feature μπορεί να εκφράζεται με έναν μόνο τρόπο ή με περισσότερους από έναν (με συνώνυμα)
    - Πχ για να αναφερθούμε στην τιμή του κινητού: {price, cost}

Βασικές Έννοιες – Opinions

- Ένα opinion passage για ένα feature f είναι ένα κομμάτι κειμένου που εκφράζει θετική ή αρνητική άποψη για το f.
- Ο opinion holder είναι αυτός που εκφράζει την άποψη.
- Μια άποψη (opinion) είναι μια θετική ή αρνητική στάση, συναίσθημα ή εκτίμηση από έναν opinion holder.
- Η πολικότητα (polarity) μιας άποψης εκφράζει αν είναι θετική, αρνητική ή ουδέτερη.

Εκτίμηση πολικότητας

- Η πολικότητα κινείται σε δύο άξονες:
  - Εκτίμηση του αν εκφράζεται κάποια άποψη ή όχι
    - SO-Polarity: Υποκειμενικό-Αντικειμενικό/Subjective-Objective
  - Εκτίμηση του τι άποψη εκφράζεται:
    - PN-Polarity: Θετικό-Αρνητικό/Positive-Negative
- Ένταση της πολικότητας
  - Πόσο θετική ή αρνητική είναι η άποψη που εκφράζεται

Εκτίμηση Πολικότητας - Λέξεις

- Για την εκτίμηση της πολικότητας μιας φράσης χρειάζεται:
  - Αναγνώριση των λέξεων που έχουν πολικότητα (opinion words).
    - Για παράδειγμα:
      - Επίθετα: {καλός, όμορφος, υπέροχος, ...}
      - Επιρρήματα: {καλά, άσχημα, ...}
      - Ουσιασικά: {σκουπίδι, ερείπιο, παράδεισος, ...}
      - Ρήματα: {μισώ, λατρεύω, ...}
      - Φράσεις και ιδιώματα: {μου κόστισε ο κούκος αηδόνι, πουλάει φούμαρα, ...}
    - Αρχικές έρευνες απέδειξαν ότι σημαντικοί δείκτες είναι τα επίθετα και επιρρήματα.
      - Τους δίνεται μεγάλη βαρύτητα στις περισσότερες προσεγγίσεις.

Εκτίμηση Πολικότητας - Σύνταξη

- Επίσης πρέπει να ληφθούν υπόψη:
  - Σύνταξη
    - Εξαρτήσεις λέξεων όταν η άποψη εκφράζεται από συνδυασμό
    - Αναγνώριση της οντότητας στην οποία αναφέρεται μια λέξη που φέρει πολικότητα
  - Άρνηση
    - Αναγνώριση της αντιστροφής στην πολικότητα

Εκτίμηση πολικότητας - Εργαλεία

- Εργαλεία που εμπλέκονται στο opinion mining:
  - Λεξικό με επισημειωμένες πολικότητες ανά λέξη
  - Μορφοσυντακτικός αναλυτής (POS tagger)
  - Συντακτικός Αναλυτής
  - Εργαλείο για επίλυση αναφορών (anaphora resolution)
  - …

SentiWordNet

- Διαθέσιμο στο:
  - http://sentiwordnet.isti.cnr.it/
- Λεξικολογική πηγή που εμπλουτίζει το WordNet
- Σε κάθε synset (έννοια, σύνολο συνωνύμων) αναθέτει τρία σκορ:
  - Θετικής πολικότητας
  - Αρνητικής πολικότητας
  - Ουδετερότητας
- Είναι διαθέσιμο σε txt μορφή.
- Κώδικας σε python:
  - http://compprag.christopherpotts.net/wordnet.html

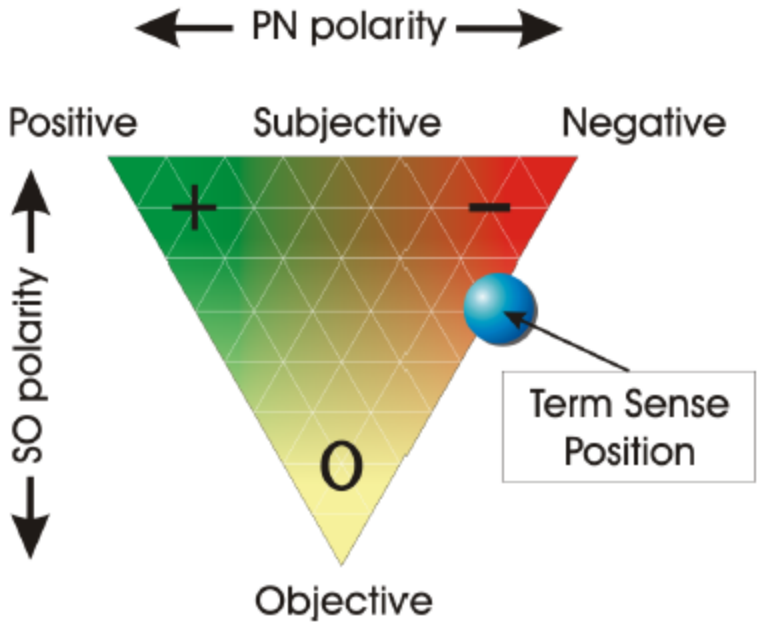

Figure 1: The graphical representation adopted by SENTIWORDNET for representing the opinion-related properties of a term sense.

SentiWordNet

- Αναγνώριση λέξεων με πολικότητα:
  - Επίλυση πολυσημίας (disambiguation)
- Αναγνώριση προσήμου και έντασης:
  - Δίνεται από τα σκορ των synsets

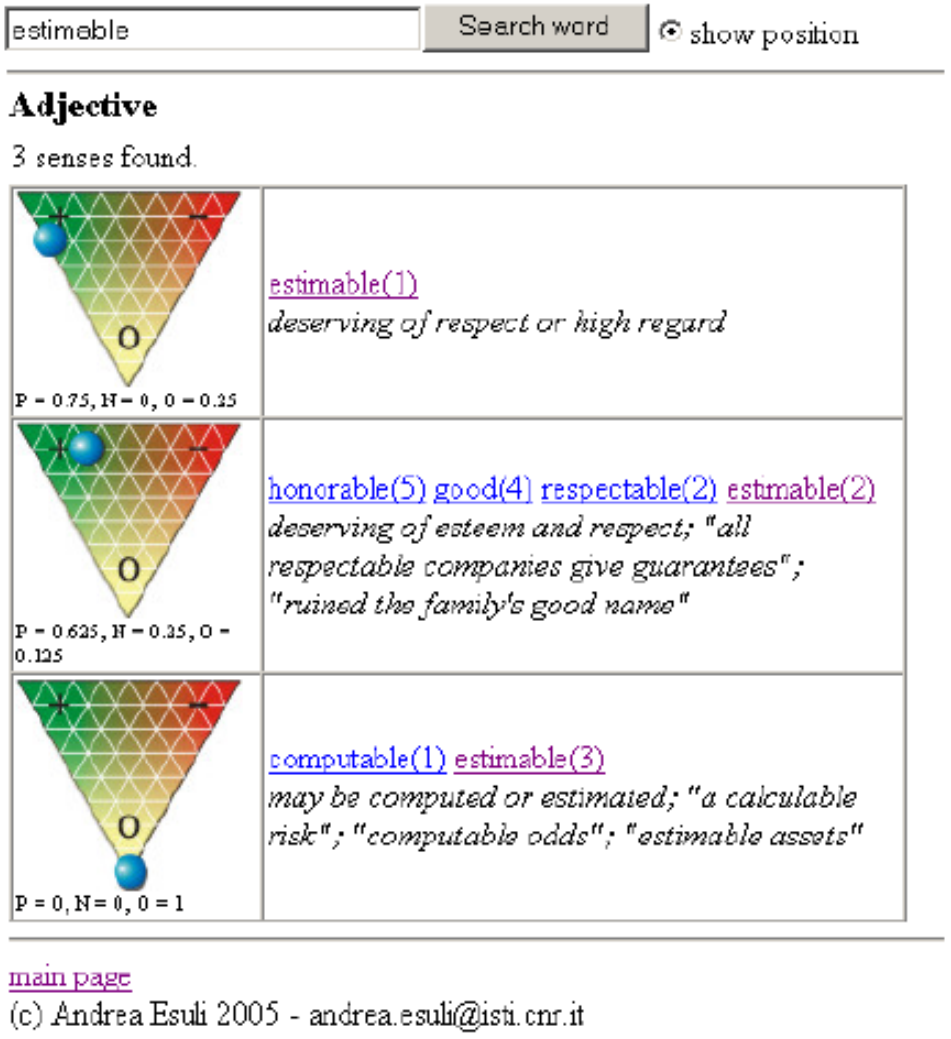

Figure 2: SENTIWORDNET visualization of the opinion-related properties of the term `estimable`.

Sentiment Classification

- Όρισμός του προβλήματος:
  - Έστω ένα σύνολο κειμένων
  - Κάθε κείμενο περιέχει σχόλια πάνω σε ένα αντικείμενο ο
  - Στόχος είναι η αναγνώριση της θετικής ή αρνητικής άποψης που εκφράζεται στο κείμενο.

Supervised Learning

- Μοντελοποίηση του προβλήματος:
  - Έστω ένα σύνολο κειμένων
  - Κάθε κείμενο περιέχει σχόλια πάνω σε ένα αντικείμενο ο
  - Στόχος είναι η ανάθεση κάθε κειμένου σε
    - Δύο κλάσεις: θετική ή αρνητική

    ή:
    - Πέντε κλάσεις: των 1-5 αστεριών
- Δεδομένα εκπαίδευσης (training set):
  - Βρίσκονται εύκολα από reviews χρηστών
  - Στην επισημείωση λαμβάνονται υπόψη τα αστεράκια
    - 1-2: negative
    - 4-5: positive

Supervised Learning

- Επιλογή χαρακτηριστικών του classification:
  - Χρήση term frequencies και tf-idf weighting:
    - Αποδεικνύονται χρήσιμα όπως στην παραδοσιακή θεματική κατηγοριοποίηση
  - Part-of-Speech tags:
    - Επιλογή των επιθέτων και επιρρημάτων
  - Opinion words and phrases:
    - Επιλογή συγκεκριμένων λέξεων και φράσεων
  - Syntactic dependency:
    - Συνυπολογισμός των συντακτικών εξαρτήσεων
  - Negation:
    - Αναγνώριση των αρνήσεων

Unsupervised Learning

- Αλγόριθμος μη-εποπτευόμενης μάθησης:
- Βήμα 1
  - Εξαγωγή των φράσεων που περιέχουν επίθετα ή επιρρήματα με βάση τα πρότυπα:

Pattern	Παράδειγμα
Επίθετο + Ουσιαστικό	It was such a *nice phone*.
Επίρρημα +Επίθετο	It was *extremely expensive*.
Επίθετο + Επίθετο	It was a *light small* silver device.
Ουσιαστικό + Επίθετο	I considered the *screen small*.
Επίρρημα + Ρηματικός τύπος	The phone was *beautifully designed*.

Unsupervised Learning

- Βήμα 2
  - Για κάθε φράση που εντοπίστηκε:
    - Υπολογισμός του pointwise mutual information (PMI)

$$PMI(word_1, word_2) = \log_2\left(\frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)}\right)$$

      - Εκφράζει τη στατιστική εξάρτηση μεταξύ λέξεων
      - Από corpus υπολογίζουμε:
        - P(word) που είναι η πιθανότητα εμφάνισης μιας λέξης
        - $P(word_1 \wedge word_2)$ είναι η πιθανότητα συνεμφάνισης δύο λέξεων
    - Υπολογισμός της πολικότητας μιας φράσης σε σχέση με τις λέξεις “excellent” ως θετική αναφορά και “poor” ως αρνητική αναφορά.
      - *SO(phrase) = PMI(phrase, “excellent”) – PMI(phrase, “poor”)*

Unsupervised Learning

- Βήμα 3
  - Με δεδομένο ένα κείμενο σχολιασμού (review) ο αλγόριθμος υπολογίζει τη μέση πολικότητα των φράσεων και
  - Ταξινομεί το κείμενο ως θετικό ή αρνητικό