Ανάκτηση Πληροφορίας: Opinion mining

Αναγνωστου αντωνιοσ 2268

Λασκαριδησ Στεφανοσ 2315

{anagnoad,laskstef}@csd.auth.gr

Περιεχόμενα

[Εισαγωγή 3](#_Toc474429473)

[Μετασχηματισμός δεδομένων 3](#_Toc474429474)

[Αλγόριθμοι μηχανικής μάθησης με επίβλεψη 4](#_Toc474429475)

[Logistic Regression 4](#_Toc474429476)

[Support Vectors Machine (SVM) 4](#_Toc474429477)

[Αλγόριθμοι μηχανικής μάθησης χωρίς επίβλεψη 4](#_Toc474429478)

[K-means 4](#_Toc474429479)

[Latent Dirichlet Allocation (LDA) 4](#_Toc474429480)

[Gaussian Mixture Model (GMM) 4](#_Toc474429481)

[Αναφορές 5](#_Toc474429482)

# Εισαγωγή

Η συγκεκριμένη εργασία εκπονήθηκε στο πλαίσιο του μαθήματος «Ανάκτηση Πληροφορίας», του 7ου εξαμήνου του Τμήματος Πληροφορικής Α.Π.Θ.

Στόχος της εργασίας είναι η μελέτη και υλοποίηση τεχνικών «Opinion mining», πάνω σε σύνολα κριτικών ταινιών από χρήστες. Κάθε κριτική μπορεί να χαρακτηριστεί είτε ως «θετική», είτε «αρνητική».

Το τεχνικό μέρος της εργασίας αναπτύχθηκε σε Apache Spark 2.0.1 και Java 1.8. Οι αλγόριθμοι που εξετάστηκαν διακρίνονται σε supervised και unsupervised και βρίσκονται εγγενώς υλοποιημένοι στο Spark.

# Μετασχηματισμός δεδομένων

Για την αναπαράσταση των δεδομένων χρησιμοποιήθηκε το διανυσματικό μοντέλο, με χρήση των μετρικών ως βάρη. Για τον διαχωρισμό των κειμένων σε όρους, χρησιμοποιούμε μία κανονική έκφραση προκειμένου να διασπαστεί το κείμενο τόσο με βάση τα κένα όσο και με βάση τα σημεία στίξης.

Για την αύξηση της αξιοπιστίας των μοντέλων, αφαιρούμε τα εξής stopwords από τους όρους:

*"i","me","my","myself","we","our","ours","ourselves","you","your","yours","yourself","yourselves","he","him","his","himself","she","her","hers","herself","it","its","itself","they","them","their","theirs","themselves","what","which","who","whom","this","that","these","those","am","is","are","was","were","be","been","being","have","has","had","having","do","does","did","doing","a","an","the","and","but","if","or","because","as","until","while","of","at","by","for","with","about","against","between","into","through","during","before","after","above","below","to","from","up","down","in","out","on","off","over","under","again","further","then","once","here","there","when","where","why","how","all","any","both","each","few","more","most","other","some","such","only","own","same","so","than","too","very","s","t","can","will","just","don","should","now","d","ll","m","o","re","ve","y"*

Επιπλέον, χρησιμοποιούμε 2-grams, προκειμένου να μειωθούν οι σημασιολογικές απώλειες.

Για την μείωση των διαστάσεων, από όρους σε έννοιες χρησιμοποιήθηκε ο αλγόριθμος Word2Vec μειώνοντας τις διαστάσεις των διανυσμάτων σε X.

# Αλγόριθμοι μηχανικής μάθησης με επίβλεψη

## Logistic Regression

## Support Vectors Machine (SVM)

# Αλγόριθμοι μηχανικής μάθησης χωρίς επίβλεψη

## K-means

## Latent Dirichlet Allocation (LDA)

## Gaussian Mixture Model (GMM)

# Αναφορές