Τεχνικές Ταξινόμησης και Συσταδοποίησης Μεγάλων Συλλογών Διαδικτυακών Δεδομένων

HOU-CS-UGP-2016-9

Παρουσίαση για την πτυχιακή εργασία

της Θεματικής ενότητας ΠΛΗ40

Μπαζάκος Κωνσταντίνος

Επιβλέπων Καθηγητής: Αναγνωστόπουλος Ιωάννης

Μέλη της Επιτροπής Κρίσης:

Αναγνωστόπουλος Χρήστος Νικόλαος

Πλαγιανάκος Βασίλειος

Επισκόπηση

- Πηγές Δεδομένων που χρησιμοποιήθηκαν (Datasets)
- Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels
- Μέθοδοι και μοντέλα ταξινόμησης
- Μέθοδοι και μοντέλα Συσταδοποίησης
- Συνεισφορά της Πτυχιακής Εργασίας

Πηγές Δεδομένων (Datasets)

Six Categories of Amazon Product Reviews

- 1,116,526 κριτικές
- 439,136 χρήστες
- 17,219 προϊόντα
- 671 χρήστες με πάνω από 50 κριτικές
- Μέσο μήκος review 667
- Περίοδος Ιουν 1999 – Ιουν 2014

Πηγές Δεδομένων (Datasets)

Amazon movie reviews dataset

- 7,911,684 κριτικές
- 889,176 χρήστες
- 253,059 προϊόντα
- 16,341 χρήστες με πάνω από 50 κριτικές
- Μέσο μήκος review 101
- Περίοδος Αυγ 1997 – Οκτ 2012

Stanford Network Analysis Project (SNAP)

Προεπεξεργασία Δεδομένων

- Μετατροπή κειμένου σε μικρά γράμματα (lowercase)
- Διαίρεση σε μεμονομένες λέξεις (tokenization)
- Αφαίρεση των stop words
- Αφαίρεση των λέξεων με μήκος μικρότερο ή ίσο με 3
- Εφαρμογή του αλγορίθμου Porter Stemmer

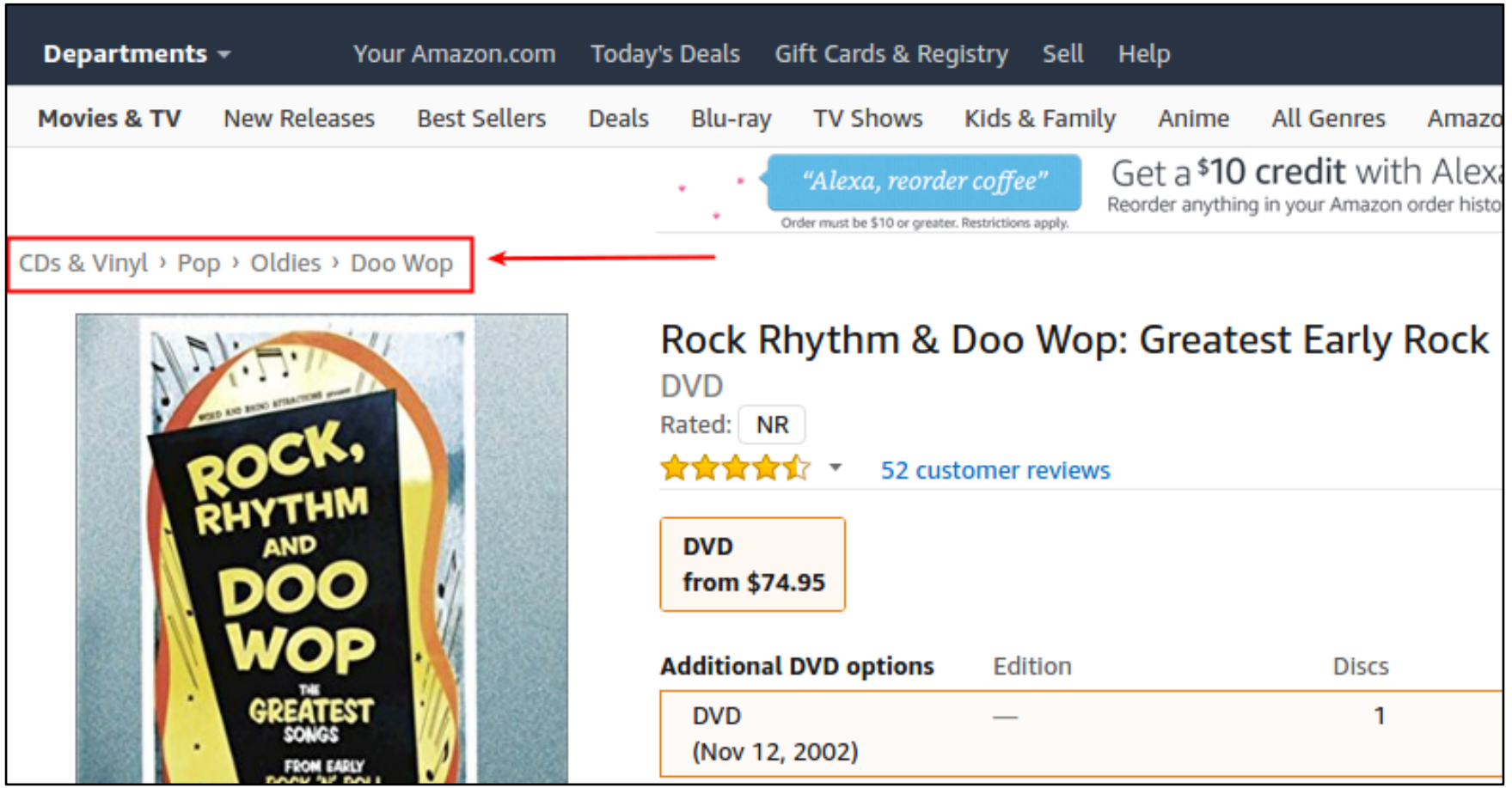
Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels

Αρχική μορφή της κάθε εγγραφής:

1. product/productId: B00006HAXW
2. review/userId: A1RSDE90N6RSZF
3. review/profileName: Joseph M. Kotow
4. review/helpfulness: 9/9
5. review/score: 5.0
6. review/time: 1042502400
7. review/summary: Pittsburgh - Home of the OLDIES
8. review/text: I have all of the doo wop DVDs...

Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels

Screenshot από προϊόν με τις κατηγορίες όπου ανήκει



Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels

Νέα μορφή της κάθε εγγραφής:

1. product/productId: B00006HAXW
2. review/userId: A1RSDE90N6RSZF
3. review/profileName: Joseph M. Kotow
4. review/helpfulness: 9/9
5. review/score: 5.0
6. review/time: 1042502400
7. review/summary: Pittsburgh - Home of the OLDIES
8. review/text: I have all of the doo wop DVDs...
9. **product/categories: ['CDs & Vinyl', 'Pop', 'Oldies', 'Doo Wop']**

Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels

- Arts, Crafts & Sewing {7}
  - Crafting {4}
    - Craft Supplies {2}
      - Craft Bells : null
      - Cutting Tools {1}
        - Scissors : null
    - Paper & Paper Crafts {3}
      - Embossing {1}
        - Embossing Folders : null
      - Paper {2}
        - Decorative Paper : null
        - Origami Paper : null
      - Punches : null
    - Sculpture Supplies {1}
      - Molding & Casting : null
    - Weaving & Spinning {3}
      - Ball Winders : null
      - Spinning Wheels : null
      - Weaving Loom Tools & Accessories : null
  - Knitting & Crochet {2}
    - Knitting Kits : null
    - Knitting Patterns : null

Ιεραρχική απεικόνηση των κατηγοριών

Μέθοδοι και μοντέλα ταξινόμησης

- Naive Bayes
- Random Forest
- Logistic Regression
- K-nearest neighbors
- Support vector machine

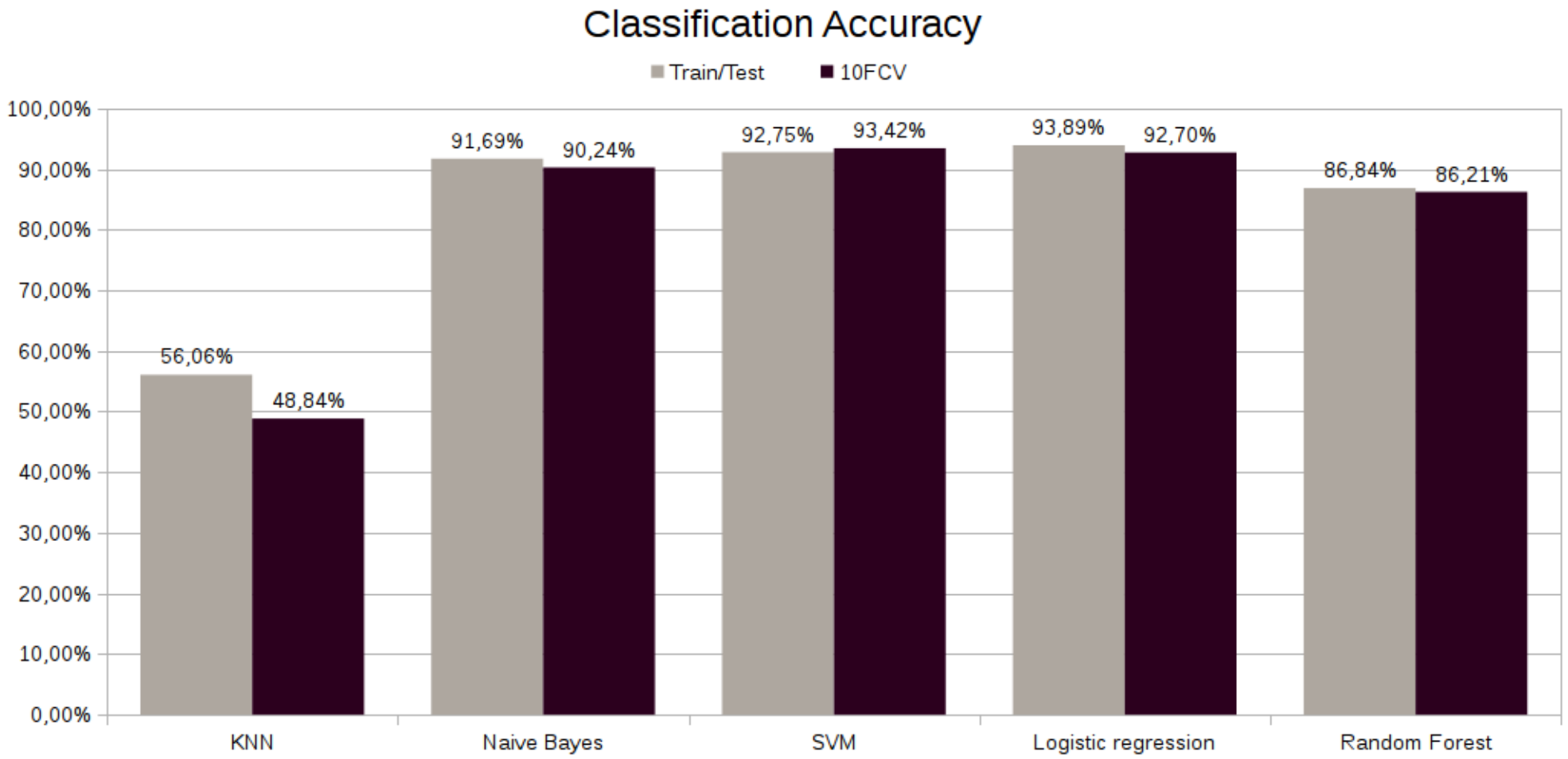
Μέθοδοι και μοντέλα ταξινόμησης

Διαδικασία υλοποίησης

- Αρχικοποίηση μοντέλου
- Εκπαίδευση με τα δεδομένα εκπαιδεύσεως (Training Set)
- Διαδικασία πρόβλεψης με είσοδο το Testing Set
- Υπολογισμός ακρίβειας των προβλέψεων
- Επιβεβαίωση αμεροληψίας του ταξινομητή βάσει k-Fold Cross Validation, με k = 10

Μέθοδοι και μοντέλα ταξινόμησης

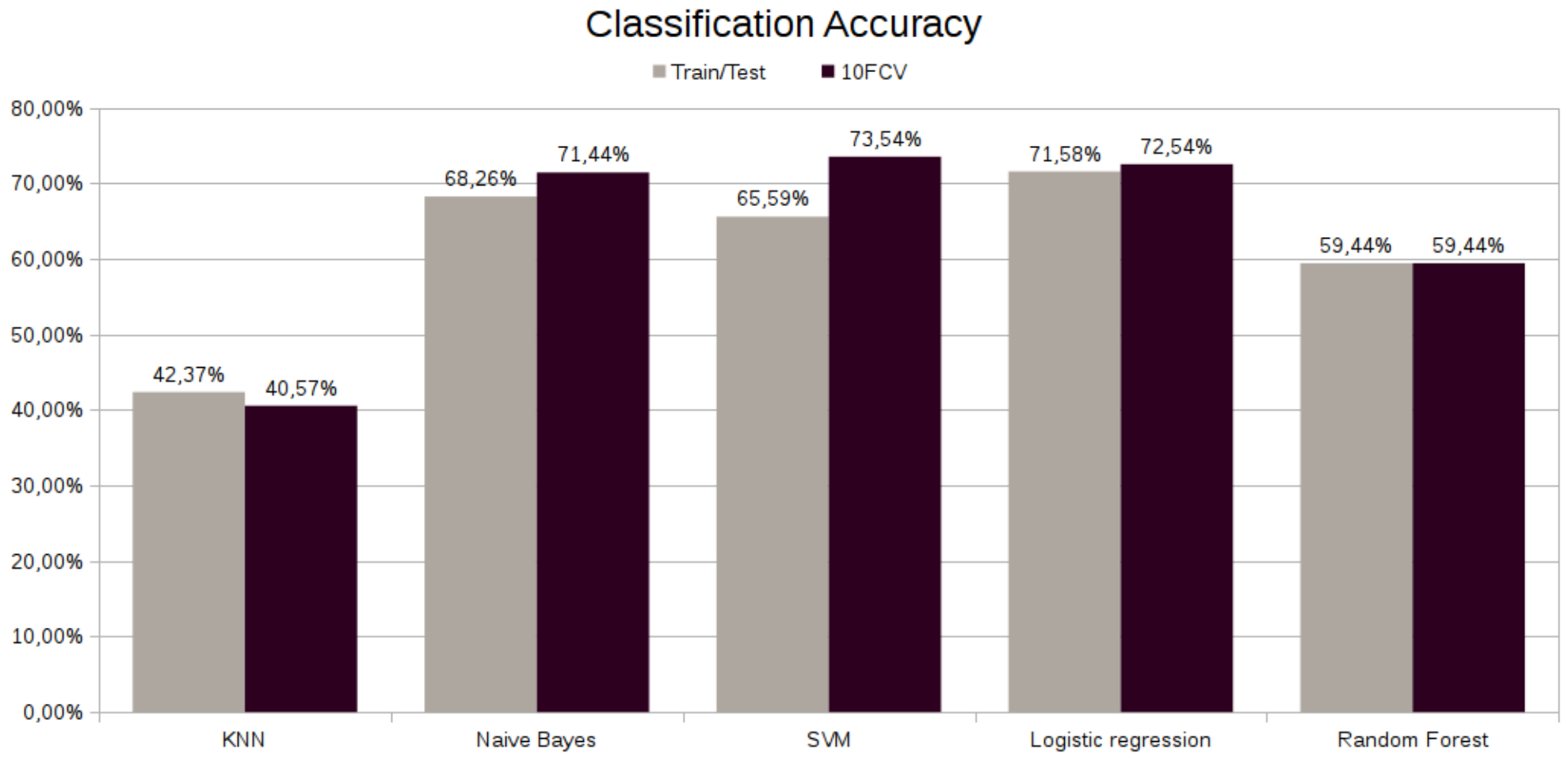
Train/Test Vs 10-fold Cross Validation



Dataset: Six Categories of Amazon Product Reviews

Μέθοδοι και μοντέλα ταξινόμησης

Train/Test Vs 10-fold Cross Validation



Dataset: Amazon Movie Reviews

Μέθοδοι και μοντέλα ταξινόμησης

Βασικές μετρικές εκτίμησης – απόδοσης

- Precision $$precision(q) = \frac{\#\ of\ \ fetched\ related\ documents}{\#\ of\ \ fetched\ documents}$$
- Recall $$recall(q) = \frac{\#\ of\ \ fetched\ related\ documents}{\#\ of\ related\ documents\ in\ collection}$$
- F_1–Score $$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Μέθοδοι και μοντέλα ταξινόμησης

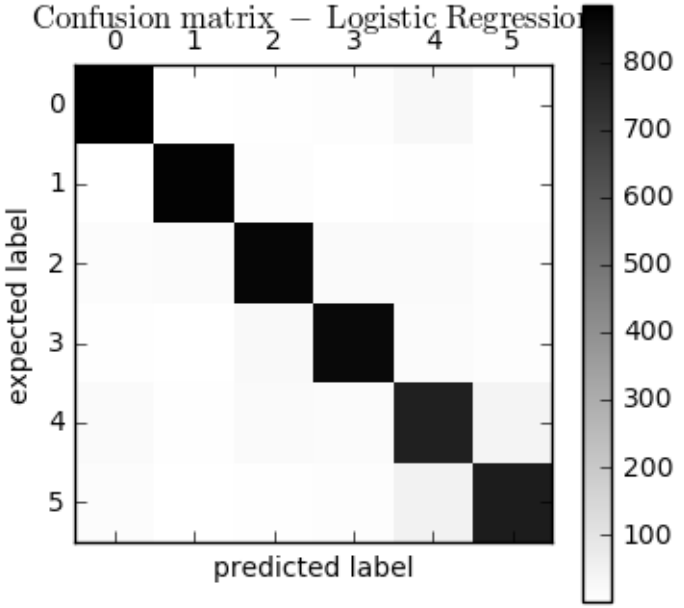
Βασικές μετρικές εκτίμησης – απόδοσης

	Naive Bayes			Logistic Regression			Random Forest			K-NN			SVM		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
0. Mobilephone	0,96	0,92	0,94	0,96	0,96	0,96	0,89	0,93	0,91	0,87	0,69	0,77	0,94	0,95	0,95
1. Cameras	0,99	0,94	0,97	0,97	0,98	0,98	0,94	0,95	0,95	0,64	0,77	0,7	0,97	0,98	0,97
2. Video Surveillance	0,87	0,97	0,91	0,95	0,92	0,94	0,89	0,88	0,88	0,86	0,26	0,41	0,93	0,92	0,92
3. Tvs	0,98	0,88	0,93	0,95	0,96	0,95	0,84	0,94	0,89	0,35	0,92	0,5	0,93	0,94	0,94
4. Tablets	0,8	0,92	0,85	0,88	0,89	0,89	0,82	0,8	0,81	0,53	0,31	0,39	0,87	0,88	0,87
5. Laptops	0,95	0,87	0,91	0,93	0,92	0,92	0,92	0,81	0,86	0,88	0,41	0,56	0,92	0,9	0,91
Average	0,93	0,92	0,92	0,94	0,94	0,94	0,88	0,89	0,88	0,69	0,56	0,56	0,93	0,93	0,93

P: Precision, R: Recall, F: F_1 – Score

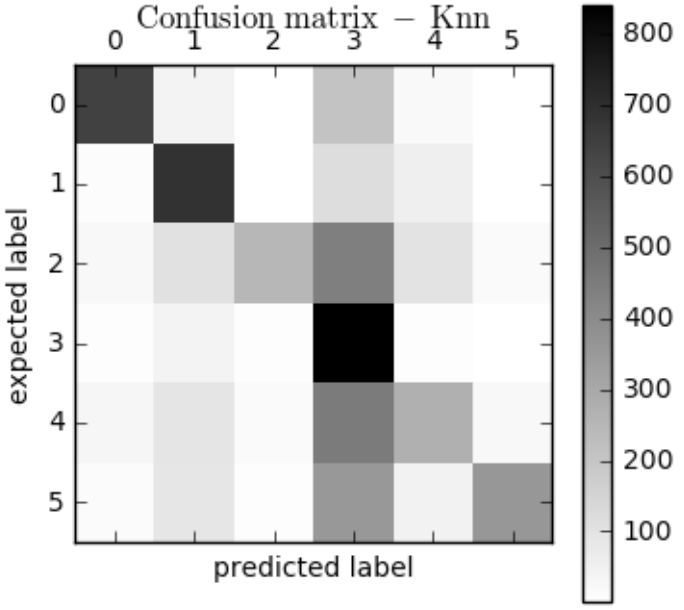
Dataset:
Six Categories of Amazon Product Reviews

Μέθοδοι και μοντέλα ταξινόμησης



Ταξινομητής: Logistic Regression

Ακρίβεια προβλέψεων: 93.89%

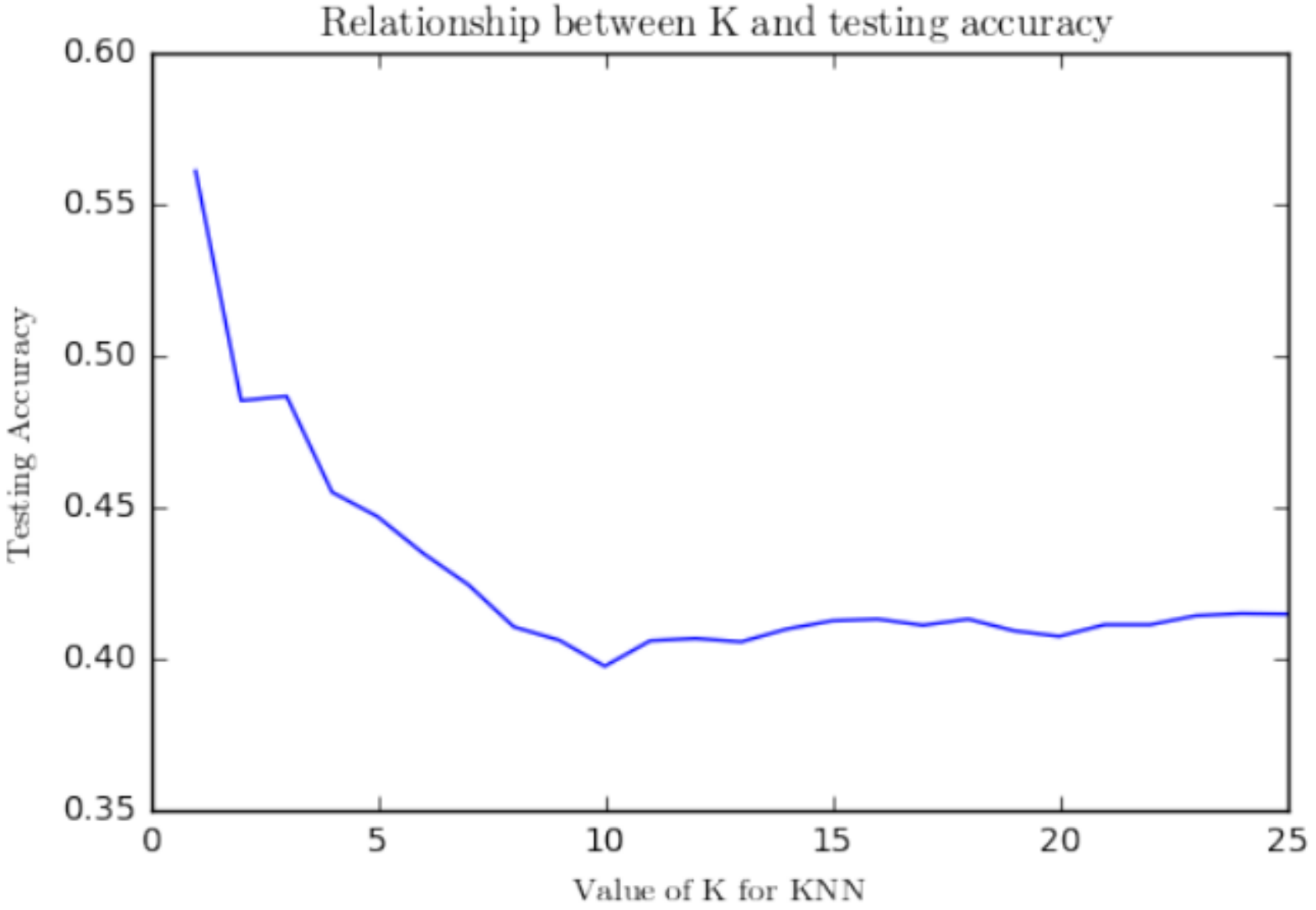


Ταξινομητής: K-nearest neighbors

Ακρίβεια προβλέψεων: 56.06%

Dataset:
Six Categories of Amazon Product Reviews

Μέθοδοι και μοντέλα ταξινόμησης



Dataset:
Six Categories of Amazon Product Reviews

Μέθοδοι και μοντέλα ταξινόμησης

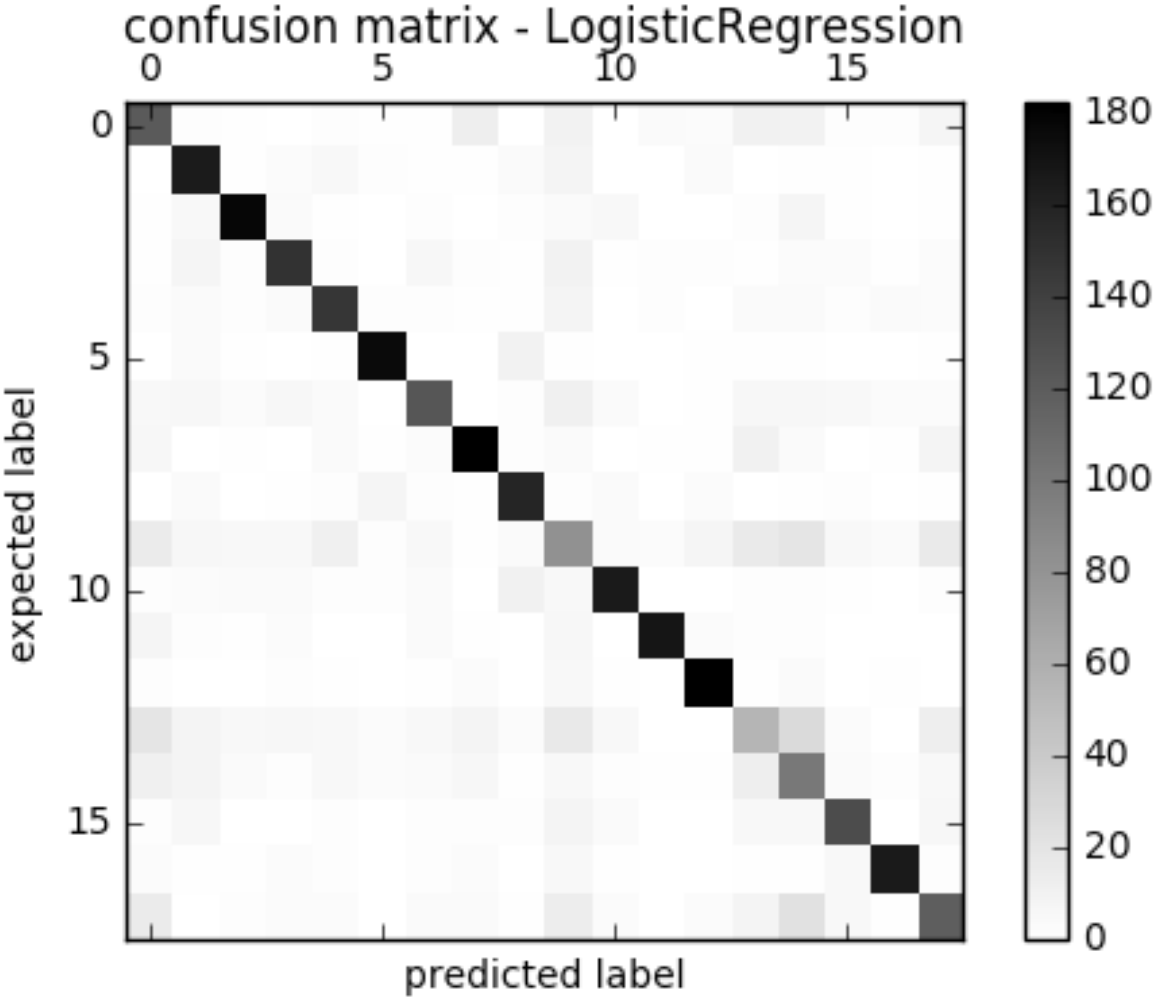
Βασικές μετρικές εκτίμησης – απόδοσης

	Naive Bayes			Logistic Regression			Random Forest			K-NN			SVM		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
0. Alternative Rock	0,38	0,81	0,51	0,56	0,64	0,6	0,29	0,44	0,35	0,38	0,27	0,31	0,4	0,56	0,47
1. Christian	0,91	0,68	0,78	0,7	0,83	0,76	0,52	0,76	0,62	0,56	0,4	0,47	0,66	0,76	0,71
2. Classical	0,74	0,89	0,81	0,84	0,83	0,84	0,7	0,75	0,72	0,8	0,31	0,45	0,8	0,78	0,79
3. R&B	0,94	0,64	0,76	0,76	0,75	0,76	0,61	0,75	0,67	0,75	0,5	0,6	0,71	0,74	0,73
4. Country	0,88	0,68	0,77	0,73	0,76	0,74	0,42	0,67	0,52	0,4	0,52	0,45	0,58	0,67	0,62
5. Children's Music	0,89	0,87	0,88	0,88	0,89	0,89	0,77	0,87	0,82	0,81	0,53	0,64	0,87	0,89	0,88
6. Jazz	0,78	0,61	0,69	0,72	0,63	0,67	0,61	0,48	0,53	0,43	0,31	0,36	0,63	0,56	0,6
7. Metal	0,89	0,64	0,74	0,79	0,79	0,79	0,71	0,65	0,67	0,73	0,42	0,54	0,72	0,71	0,71
8. Special Interest	0,81	0,76	0,78	0,78	0,84	0,81	0,69	0,73	0,71	0,91	0,33	0,48	0,78	0,8	0,79
9. Pop	0,43	0,29	0,35	0,39	0,38	0,39	0,25	0,2	0,22	0,14	0,21	0,17	0,3	0,32	0,31
10. New Age	0,93	0,61	0,74	0,83	0,79	0,81	0,79	0,63	0,7	0,71	0,56	0,63	0,82	0,73	0,77
11. Dance & Electronic	0,88	0,85	0,86	0,93	0,84	0,88	0,92	0,77	0,84	0,82	0,67	0,74	0,89	0,78	0,83
12. Rap & Hip-Hop	0,92	0,85	0,88	0,86	0,9	0,88	0,81	0,84	0,83	0,15	0,8	0,26	0,86	0,86	0,86
13. World Music	0,29	0,43	0,34	0,39	0,3	0,34	0,27	0,13	0,18	0,27	0,13	0,18	0,35	0,28	0,31
14. Rock	0,41	0,61	0,49	0,45	0,55	0,49	0,36	0,34	0,35	0,29	0,23	0,26	0,42	0,45	0,43
15. Blues	0,78	0,69	0,73	0,77	0,73	0,75	0,65	0,6	0,62	0,64	0,48	0,55	0,72	0,66	0,69
16. Folk	0,95	0,73	0,83	0,9	0,85	0,87	0,91	0,76	0,83	0,83	0,67	0,74	0,89	0,8	0,84
17. Classic Rock	0,5	0,67	0,57	0,61	0,57	0,59	0,5	0,33	0,4	0,43	0,25	0,32	0,53	0,44	0,48
Average	0,74	0,68	0,70	0,72	0,72	0,71	0,60	0,59	0,59	0,56	0,42	0,45	0,66	0,66	0,66

P: Precision, R: Recall, F: F_1 – Score

Dataset: Amazon Movie Reviews

Μέθοδοι και μοντέλα ταξινόμησης

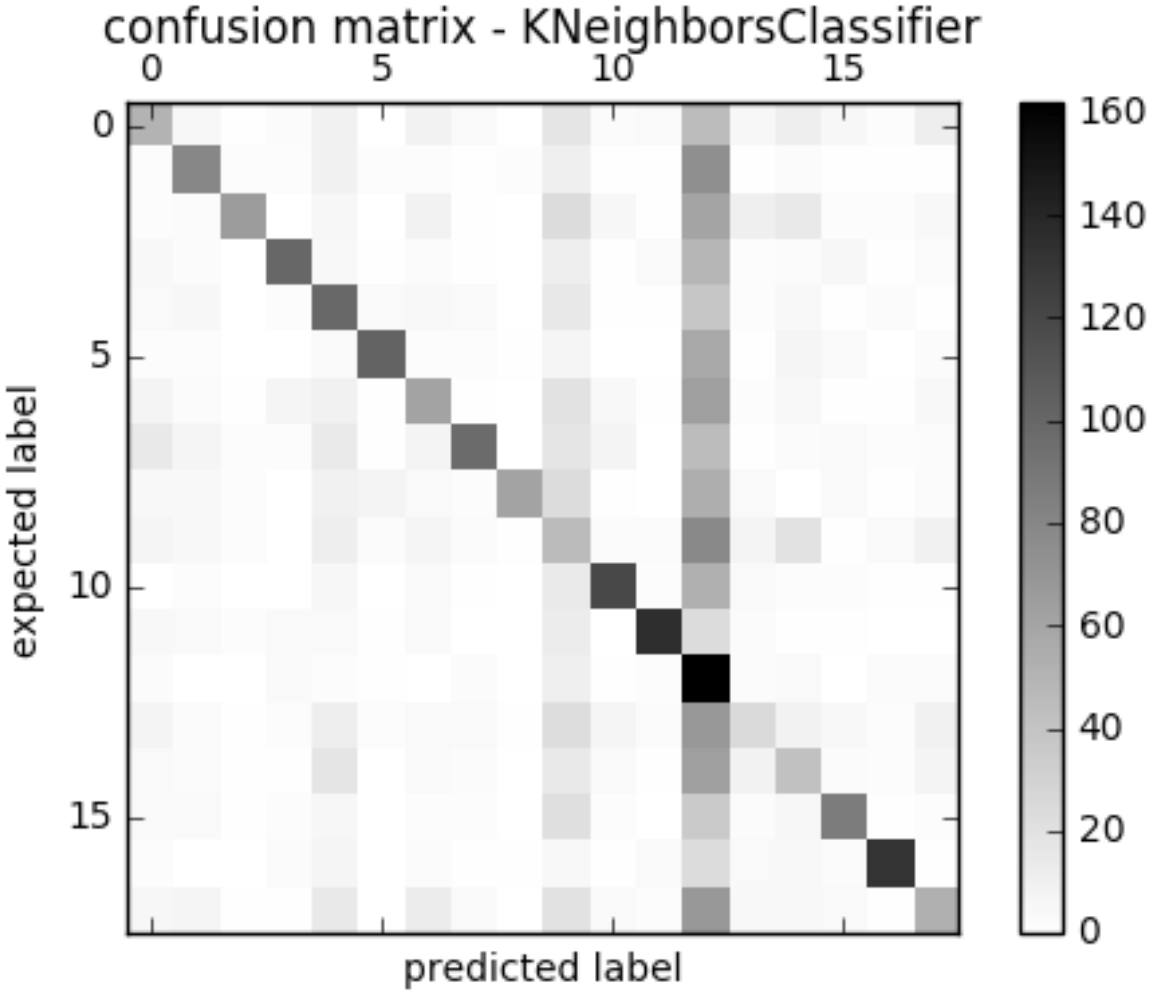


Ταξινομητής:
Logistic Regression

Ακρίβεια προβλέψεων:
71.58%

Dataset:
Amazon Movie Reviews

Μέθοδοι και μοντέλα ταξινόμησης

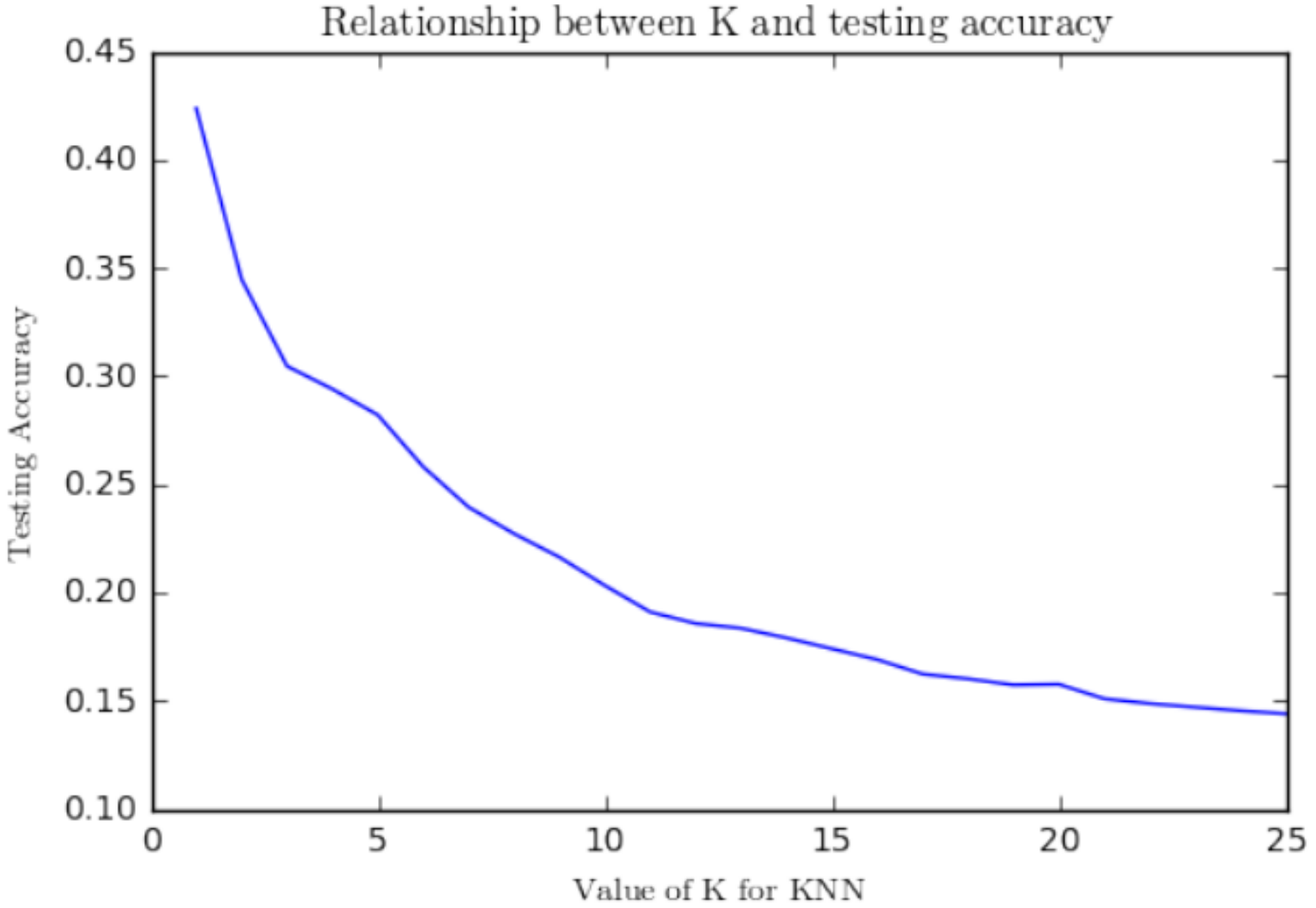


Ταξινομητής:
K-nearest neighbors

Ακρίβεια προβλέψεων:
42.37%

Dataset:
Amazon Movie Reviews

Μέθοδοι και μοντέλα ταξινόμησης



Dataset:
Amazon Movie Reviews

Μέθοδοι και μοντέλα Συσταδοποίησης

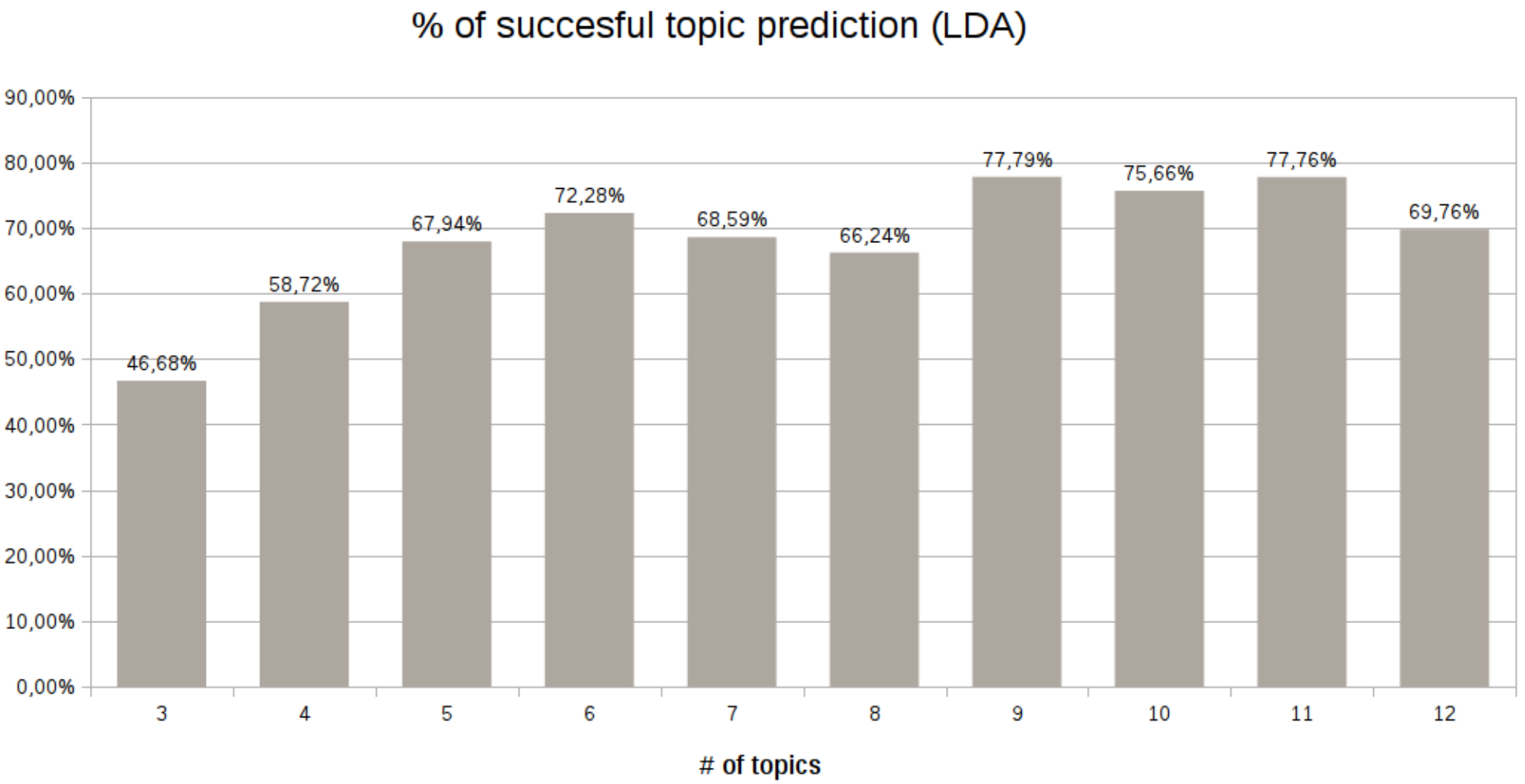
- Latent Dirichlet Allocation (LDA)
  - Topic model (# of topics [n/2, 2*n], n=6)
- Πίνακες/Μάσκες Συνάφειας
  - Βάσει εσφαλμένων ταξινομήσεων

Μέθοδοι και μοντέλα Συσταδοποίησης

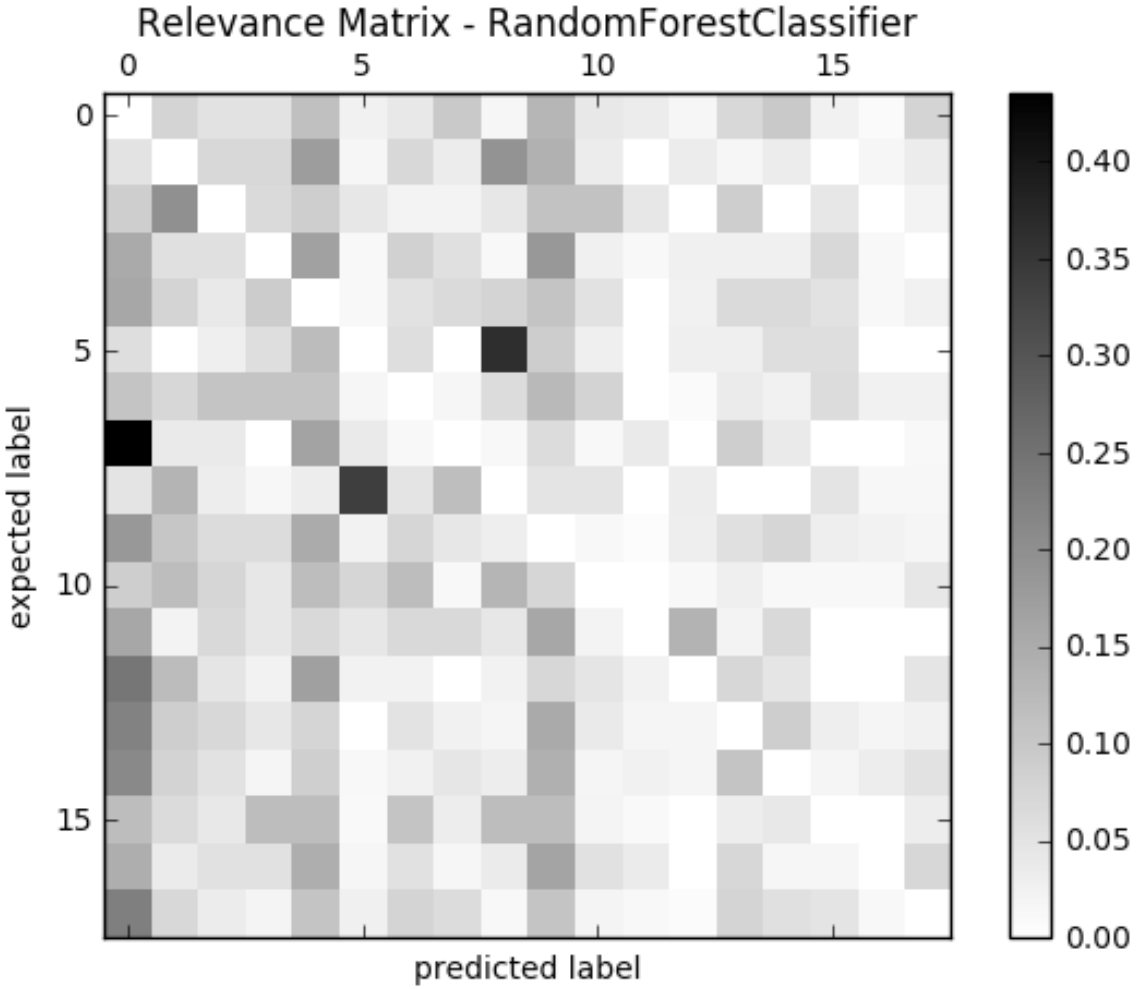
Εξαγωγή topics και προβλήματα (LDA)

- Λυσαρεστημένοι χρήστες
- Ευχαριστημένοι χρήστες
- Επιστροφές προϊόντων
- Γλώσσες όπως Κινέζικα, Ισπανικά κ.λπ.
- Μονολεκτικά reviews

Μέθοδοι και μοντέλα Συσταδοποίησης



Μέθοδοι και μοντέλα Συσταδοποίησης



Μάσκα συνάφειας βάσει ταξινομητή:

Random Forest

Dataset:
Amazon Movie Reviews

Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels

Η εμπλουτισμένη αυτή πηγή δεδομένων (όπως και ο κώδικας της ΠΕ) είναι διαθέσιμα στην κοινότητα μέσω git repository στο GitHub.com. Επίσης φιλοξενείται ήδη στα Datasets της Kaggle.com

Βρίσκεται σε εξέλιξη η ανάρτηση του στην επίσημη ιστοσελίδα του του Stanford University (SNAP)

https://github.com/bazakoskon/Classification-clustering-Thesis

https://bazakoskon.github.io/labels-on-Amazon-movie-reviews-dataset

kaggle

https://kaggle.com/thebuzz/ground-truth-labels-amazon-movie-reviews-dataset

Τελική αποτίμηση
Μελλοντικές επεκτάσεις

- Πηγές Δεδομένων που χρησιμοποιήθηκαν (Datasets)
- Εμπλουτισμός του “Amazon Movie Reviews Dataset” με ground truth labels
- Μέθοδοι και μοντέλα Ταξινόμησης και Συσταδοποίησης
- Μελλοντικές επεκτάσεις
  - Πιθανή παρουσίαση της εργασίας σε κάποιο συνέδριο
  - Υλοποίηση συστήματος multi-label classification στο εμπλουτισμένο dataset

Τέλος παρουσίασης

Σας ευχαριστώ!