MAI4CAREU

Επεξεργασία Φυσικής Γλώσσας

Text Classification

Δημήτρης Πασχαλίδης

Τμήμα Επιστήμης Υπολογιστών

Πανεπιστήμιο Κύπρου

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Classification

- Μια αντιστοίχιση h από δεδομένα εισόδου x (από instance space X) σε μια ετικέτα y (από categorical space Y).
  - X = σύνολο όλων documents
  - Y = $\{y_1, y_2, \ldots, y_k\}$
  - x = ένα μονό document
  - y = ετικέτα εξόδου για x

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Movie Reviews

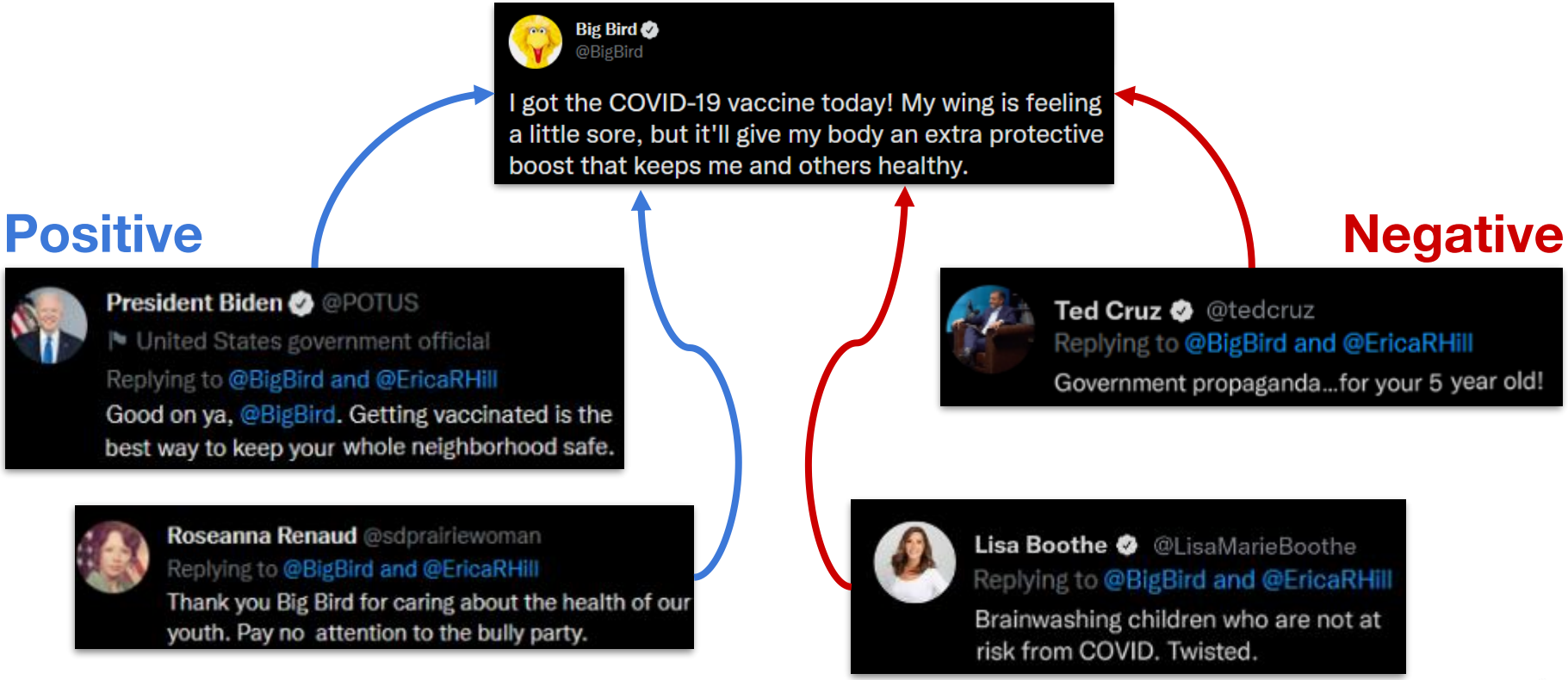
- Y = {Positive, Negative, Neutral}

- x_1 = “This film blew me away, exciting, fast paced, surprisingly gritty, and genuinely had an awesome story…”
- y_1 = **Positive**

- x2 = “... This boring 3 hour movie had no proper plot whatsoever but just a bunch of random scenes and stupid dialogues ...”
- y_2 = **Negative**

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Opinion Mining



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συγγραφέας: Αρσενικός or Θηλυκός ?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimori, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, vol. 23, p. 321-346

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Είναι αυτό SPAM ?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

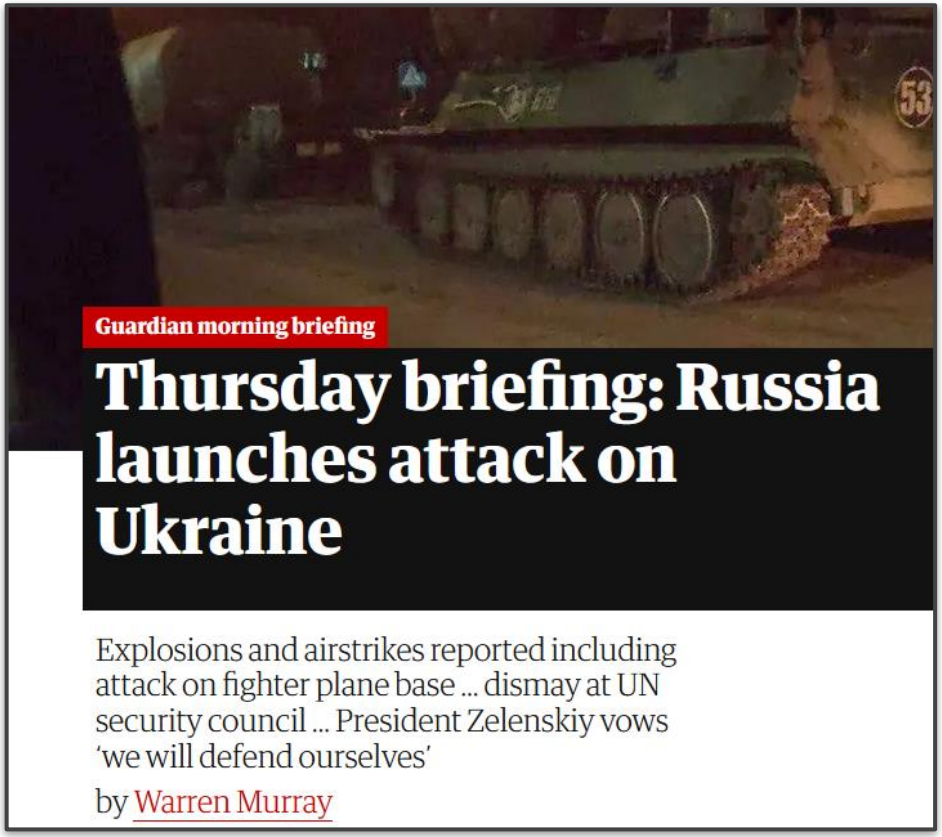
Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Θέμα άρθρων



- ❑ Politics
- ❑ Sports
- ❑ Lifestyle
- ❑ Gossip
- ❑ Health
- ❑ …

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Text Classification Πρόβλημα

Δοσμένου ενός κειμένου $w = (w_1, w_2, \ldots, w_T) \in V^*$,
πρόβλεψη ετικέτας $y \in Y$

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Text Classification Εφαρμογές

Task	X	Y
Language Identification	Text	{ English, Greek, Mandarin, … }
SPAM Classification	Email	{ SPAM, not SPAM }
Authorship Attribution	Text	{ J.K. Rowling, George R. R. Martin, … }
Genre Classification	Novel	{ Detective, Romance, Fantasy, … }
Sentiment Classification	Text	{ Positive, Negative, Neutral, Mixed }
Hate-speech Classification	Text	{ Toxic, Sexist, Hate, Racism, … }
Veracity Classification	Text	{ Fake, Real, Reliable, Unreliable }

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Classification Μέθοδοι:
Rule-based Classification

- Κανόνες που βασίζονται σε μοτίβα, συνδυασμούς λέξεων ή άλλα χαρακτηριστικά.
  - SPAM / UNRELIABLE:
    - Black-listed πηγές e.g. www.realnews.124.info
- Εάν οι κανόνες βελτιωθούν από τους ειδικούς → υψηλές βαθμολογίες ακρίβειας.

Δημιουργία / Διατήρηση / Φιλτράρισμα κανόνων

→ <u>Time Consuming & Resource Demanding</u>

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Classification Μέθοδοι :
Supervised Machine Learning

- Δεδομένα εισόδου:
  - ένα έγγραφο ***d***
  - ένα σταθερό σύνολο κλάσεων / ετικετών $C = \{ c_1, c_2, \ldots, c_k \}$
  - ένα εκπαιδευτικό σύνολο annotated έγγραφα

    $m = \{ (d_1, c_1, \ldots, (d_m, c_m) \}$
- Έξοδος:
  - ένας μαθημένος classifier $y : d \rightarrow c$

Examples of Classifiers:
- Naive Bayes
- Logistic Regression
- Support Vector Machines
- k-Nearest Neighbors

…

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

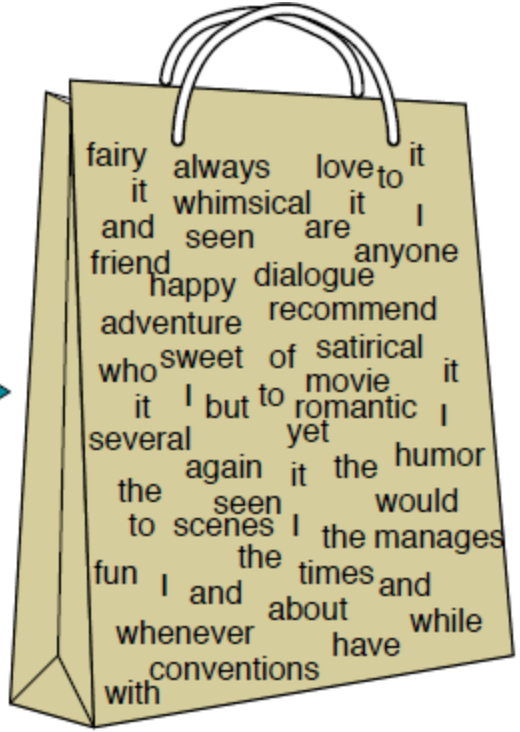
Text Vectorization

- Οι υπολογιστές δεν κατανοούν το κείμενο, τη σύνταξη και τη σχέση μεταξύ των λέξεων → **Χρειάζεστε έναν τρόπο αναπαράστασης λέξεων με αριθμούς**.
- Text vectorization: Η διαδικασία μετατροπής κειμένου σε αριθμητική αναπαράσταση.
  - Δημοφιλής βασική αναπαράσταση κειμένου: **Bag of Words** (**BoW**)

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bag-of-Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bag-of-Words Representation

- Ο bag-of-words είναι μια αναπαράσταση σταθερού μήκους, η οποία αποτελείται από ένα διάνυσμα καταμέτρησης λέξεων:

s = "*It was the best of times*, *it was the words of times*"

0 1 2 2 0

x = [*aardvark*, …, *best*, …, *it*, …, *of*, …, *zyther*]

- Το μήκος του x ισούται με το μέγεθος του λεξιλογίου
- Για κάθε x, μπορεί να υπάρχουν πολλά w, ανάλογα με τη σειρά των λέξεων

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Linear Classification επί BoW

- Εστω $\psi(x, y)$ χαρακτηρίζει την συμβατότητα του bow x και ετικέτας y, τότε $\hat{y} = \underset{y}{\operatorname{argmax}}\, \psi(x, y)$
- Σε ενα linear classifier, Αυτή η συνάρτηση βαθμολόγησης έχει τη μορφή $\psi(x, y) = \theta \cdot f(x, y) = \sum_{j=1} \theta_j \cdot f_j(x, y)$
  - ώπου θ είναι ένα διάνυσμα **weights** / **coefficients** και f είναι ένα **feature function**.

Learning problem: έυρεση σωστών weights θ, υποθέτοντας $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πιθανοτικό Classification

- Naive Bayes είναι πιθανοτικός classifier με βάση την ακόλουθη στρατηγική:
  - Ορισμός μοντέλου πιθανότητας $p(x, y)$
  - Εκτιμήστε τις παραμέτρους του μοντέλου πιθανότητας με τη μέγιστη πιθανότητα στο training dataset.
- **Bayes Rule** για έννραφο d και κλάση / ετικέτα c :

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP είναι “maximum a posteriori” = Πιο πιθανή κλάση

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

P(***d***) καταργείται επειδή είναι περιττή. Για ***d***, η πιθανότητά του είναι η ίδια για όλα τα ***c***.

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes Classifier

Likelihood

Prior

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Έγγραφο *d* εκπροσωπείτε ως features $x_1 \ldots x_n$

$O(|X|^n \times |C|)$ parameters → Εάν έχουμε πολλά χαρακτηριστικά, τότε χρειαζόμαστε μεγάλο αριθμό παραδειγμάτων training.

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Multinomial Naive Bayes
Independence Assumptions

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **BoW Assumption**: **Η τοποθέτηση λέξεων δεν έχει σημασία**.
- **Conditional Independence**: Feature πιθανότητες $P(x_i \mid c_j)$ είναι ανεξάρτητες δεδομένης της κατηγορίας / ετικέτας c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Multinomial Naive Bayes

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c_j) \prod_{x \in X} P(x \mid c)$$

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\, P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

Για text classification:

Θέσεις = όλες οι θέσεις λέξεων στο έγγραφο τεστ.

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προβλήματα με Probability Product

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

- **Ο πολλαπλασιασμός πολλών πιθανοτήτων μπορεί να οδηγήσει σε αριθμητικό underflow**.
  - 0.0006 x 0.0007 x 0.0009 x 0.01 x 0.5 x 0.000008 …
- Λύση: Χρήση logs → log(a x b) = log(a) + log(b)

$$c_{\mathrm{NB}} = \underset{c_j \in C}{\operatorname{argmax}} \left[\log P(c_j) + \sum_{i \in \mathrm{positions}} \log P(x_i | c_j) \right]$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Learning του Naive Bayes

MAI4CAREU

- Learning του Naive Bayes:
  - Χρήση του Laplace (add-1) smoothing για την αντιμετώπιση μηδενικών πιθανοτήτων.

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)} = \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

1. Απόσπασμα λεξιλογίου V από corpus

2. Υπολογίσμος $P(c_j)$ όρων:

$\forall\ c_j \in C \rightarrow docs_j$ = docs από c_j

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

3. Υπολογίσμος $P(w_k \mid c_j)$ όρων :

$text_j$ = συγχώνευση $docs_j$

$\forall\ w_k \in V \rightarrow n_k = occur(w_k, text_j)$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Sentiment Analysis
με Naive Bayes

Document	Category
Just plain boring	Negative
Entirely predictable and lacks energy	Negative
No surprises and very few laughs	Negative
Very powerful	Positive
The most fun film of the summer	Positive
Predictable with no fun	?

Training

Test

1. + / - Πιθανότητες

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

P(-) = 3 / 5

P(+) = 2 / 5

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Sentiment Analysis
με Naive Bayes

Document	Category
Just plain boring	Negative
Entirely predictable and lacks energy	Negative
No surprises and very few laughs	Negative
Very powerful	Positive
The most fun film of the summer	Positive
Predictable ~~with~~ no fun	?

Training

Test

1. + / - Πιθανότητες

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

P(-) = 3 / 5

P(+) = 2 / 5

2. Clean Texts

Remove "*with*".

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

3. Υπολογισμός Training Likelihoods

Document	Category
Just plain boring	Negative
Entirely predictable and lacks energy	Negative
No surprises and very few laughs	Negative
Very powerful	Positive
The most fun film of the summer	Positive
Predictable ~~with~~ no fun	?

$$p(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20}$$

$$P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“fun”}|-) = \frac{0+1}{14+20} \quad P(\text{“fun”}|+) = \frac{1+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \quad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

4. Test Set Scoring

Document	Category
Just plain boring	Negative
Entirely predictable and lacks energy	Negative
No surprises and very few laughs	Negative
Very powerful	Positive
The most fun film of the summer	Positive
Predictable ~~with~~ no fun	?

$$P(-)P(S|\ -)\ =\ \frac{3}{5}\times\frac{2\times2\times1}{34^3}=6.1\times10^{-5}$$

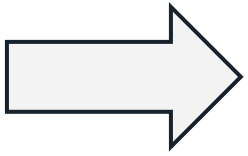
$$P(+)P(S|\ +)\ =\ \frac{2}{5}\times\frac{1\times1\times2}{29^3}=3.2\times10^{-5}$$

Το έγγραφο δοκιμής είναι πιο πιθανό να είναι **NEGATIVE** class

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes και LM

- Naive Bayes classifier μπορεί να χρησιμοποιήσει διαφορετικά features:
  - URLs, emails, dictionaries, numerical και network features.
- Ωστόσο, αν χρησιμοποιούμε μόνο word features, και εξετάστε το whole corpus (not just a subset of words)



Naive Bayes είναι παρόμοιο με Language Modeling από τις προηγούμενες διαλέξεις.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes και LM

- Αντιστοίχιση κάθε λέξης: $P(word \mid c)$
- Ανάθεση κάθε πρότασης: $P(s \mid c) = \prod P(word \mid c)$

Class POSITIVE	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.05	0.01	0.1

$$P(s \mid POSITIVE) = 0.0000005$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes ως LM

❑ Ποια τάξη αποδίδει τη μεγαλύτερη πιθανότητα σε *s*?

Class POSITIVE	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Class NEGATIVE	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.05	0.01	0.1
0.2	0.001	0.01	0.005	0.1

$$P(s \mid POSITIVE) = 0.0000005$$

$$P(s \mid NEGATIVE) = 0.0000001$$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Naive Bayes ως LM

❑ Ποια τάξη αποδίδει τη μεγαλύτερη πιθανότητα σε *s*?

Class POSITIVE	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Class NEGATIVE	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.05	0.01	0.1
0.2	0.001	0.01	0.005	0.1

P(*s* | *POSITIVE*) = 0.0000005

P(*s* | *NEGATIVE*) = 0.0000001

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Άλλοι Classifiers

- Naive Bayes ονομάζεται έτσι λόγω του:
  - Αγνοώντας αφελώς τις εξαρτήσεις μεταξύ των λέξεων και τη σειρά τους, αντιμετωπίζοντας κάθε λέξη ως εξίσου ενημερωτική.
  - Οι feature πιθανότητες P(xi | cj) είναι ανεξάρτητες δεδομένης της κατηγορίας / ετικέτας c.

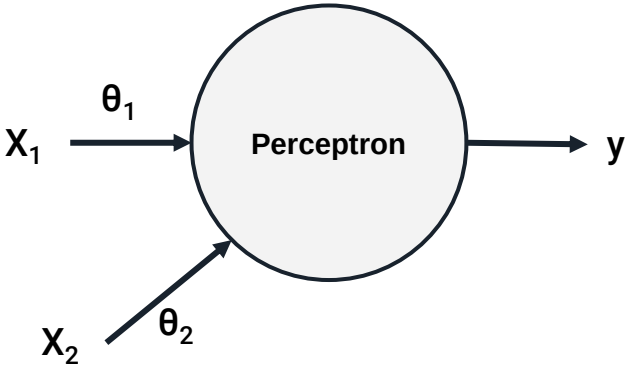
$$P(x_1,\ldots,x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

    - Discriminative classifiers αποφύγουν αυτό το ζήτημα, χωρίς να μοντελοποιουν τη "παραγωγική" πιθανότητα $p(x)$

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Perceptron Classifier

- Υπόθεση που βασίζεται σε σφάλματα και όχι στην υπόθεση ανεξαρτησίας.
- Perceptron Learning Κανόνας:
  - Εκτελέστε το τρέχον classifier σε μια παρουσία training data, υπολογίζοντας $y^{\wedge} = \text{argmax}_y \psi(x^{(i)}, y)$
  - Εάν η πρόβλεψη είναι λανθασμένη:
    - Αυξήστε τα βάρη για το features της πραγματικής ετικέτας.
    - Μειώστε τα βάρη για το features της προβλεπόμενης ετικέτας. $\theta \leftarrow \theta + f(x^{(i)}, y^{(i)}) - f(x^{(i)}, y^{\wedge})$
  - Επαναλάβετε μέχρι τα training instances είναι σωστά classified, ή εξαντλείται ο χρόνος.



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Perceptron Classifier

Algorithm 3 Perceptron learning algorithm

procedure PERCEPTRON($\boldsymbol{x}^{(1:N)}, y^{(1:N)}$)
  $t \leftarrow 0$
  $\boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}$
  repeat
    $t \leftarrow t + 1$
    Select an instance i
    $\hat{y} \leftarrow \text{argmax}_y \, \boldsymbol{\theta}^{(t-1)} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)$
    if $\hat{y} \neq y^{(i)}$ **then**
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \boldsymbol{f}(\boldsymbol{x}^{(i)}, \hat{y})$
    else
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$
  until tired
  return $\boldsymbol{\theta}^{(t)}$

- Objective: Ελαχιστοποίηση loss function στα βάρη.
- Loss functions θα πρέπει να είναι:
  - Ένας καλός πληρεξούσιος της ακρίβειας των classifiers.
  - Εύκολη βελτιστοποίηση.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Περισσότερα για Sentiment Analysis

- Αντιμετώπιση της **Άρνησης**

  I really like this movie.
  - I really **don't** like this movie.

  - **Doesn't** let us get bored.
  - **Don't** miss it!

- Η άρνηση αντιστρέφει το νόημα του συναισθήματος.
- Απλή μέθοδος baseline: Δημιουργία νέων λέξεων / tokens με προσθήκη **NOT_** μεταξύ εμφάνισης άρνησης και στίξης.

```
Didn't like this movie.
Didn't NOT_like NOT_this NOT_movie.
```

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Περισσότερα για Sentiment Analysis

- Sentiment Lexicons
  - Αν δεν έχετε αρκετά training / labeled data χρησιμοποιήστε προκατασκευασμένες λίστες λέξεων, με το όνομα **Lexicons**.
- **MPQA Subjectivity Cues**: 2718 POSITIVE και 4912 NEGATIVE λέξεις / φράσεις με ένταση (ισχυρές / αδύναμες)
- **General Inquirer**: 1915 POSITIVE και 2291 NEGATIVE, με annotations Ισχυρών vs Αδύναμων, Ενεργητικών vs Παθητικών, Υπερεκτιμημένων vs. Υποτιμημένων και Ηδονής, Πόνου, Αρετής, Κακίας, Κινήτρων, Γνωστικού Προσανατολισμού, etc

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Περισσότερα για Sentiment Analysis

- **Opinion Lexicon**: 2006 POSITIVE και 4789 NEGATIVE λέξεις / φράσεις με ένταση (δυνατές/αδύναμες)
- **AFINN**: AFINN lexicon είναι μια λίστα αγγλικών όρων που βαθμολογούνται για το σθένος μεταξύ -5 (NEGATIVE) και +5 (POSITIVE)
- **Loughran-McDonald**: Αγγλικό sentiment lexicon δημιουργήθηκε για χρήση με οικονομικά έγγραφα → **Financial Sentiment Analysis**

❑ Χρησιμοποιώντας Lexicons

- Διάνυσμα των μετρήσεων κάθε φορά που εμφανίζεται μια λέξη από το λεξικό.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Text Classification Αξιολόγηση

	Correct	Not Correct
Selected	TP	FP
Not Selected	FN	TN

Confusion Matrix

- **Correct**: Τι γνωρίζουμε ότι είναι σωστό στο dataset.
- **Not Correct**: Τι γνωρίζουμε ότι δεν είναι σωστό στο dataset.
- **Selected**: Τι επέστρεψε ο classifier (αποτέλεσμα).
- **Not Selected**: Τι δεν επέστρεψε ο classifier.
- **TP (True Positive)**: Τι επέστρεψε ο classifier και είναι ορθό.
- **FP (False Positive)**: Τι επέστρεψε ο classifier και δεν είναι ορθό.
- **FN (False Negative)**: Τί δεν επέστρεψε ο classifier αλλα ήταν ορθό.
- **TN (True Negative)**: Τί δεν επέστρεψε ο classifier και δεν ήταν ορθό.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Precision, Recall, και F Measure

- **Precision**: % των στοιχείων που πεστρεψε και είναι σωστά
- **Recall**: % σωστών στοιχείων που επέστρεψε
- **F Measure**: Συνδυασμένο μέτρο που αξιολογεί την Precision και Recall tradeoff (Weighted Harmonic Mean)

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

F = 2PR

Συνήθως χρησιμοποιείται ως ισορροπημένη F1 με β=1 (or α=½)

	Correct	Not Correct
Selected	TP	FP
Not Selected	FN	TN

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Multi-Class Αξιολόγηση

- **Precision**: Ποσοστό εγγράφων της κλάσης *i* και αφορούν την τάξη *i*:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

- **Recall**: Ποσοστό εγγραφών της κλάσης i που έγιναν classified σωστά:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

- **Accuracy**: Ποσοστό εγγραφών που έγιναν classified σωστά:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Micro- και Macro- Averaging

- Για multi-class classification, συνδυάστε μετρήσεις απόδοσης ανά κατηγορία με micro- και macro- averaging.

- **Macro-Averaging**: Υπολογισμός απόδοσης για κάθε κλάση και μετά average.
- **Micro-Averaging**: Υπολογισμός contingency table για όλες τις τάξεις και αξιολόγηση

Class 1

	Truth: YES	Truth: NO
CLF: YES	10	10
CLF: NO	10	970

Class 2

	Truth: YES	Truth: NO
CLF: YES	90	10
CLF: NO	10	890

Class 3

	Truth: YES	Truth: NO
CLF: YES	100	20
CLF: NO	20	1860

- **Macro-Averaged Precision:** (0.5 + 0.9) / 2 = 0.7
- **Micro-Averaged Precision:** 100 / 120 = 0.83
- Micro-Averaged Score is dominated by score on common classes.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Περίληψη

- Τύποι Classification:
  - Rule-based Classification
  - Supervised Classification
- Εισαγωγή στο Text Vectorization → Είσοδος σε ML models.
  - Παράδειγμα vectorization: Bag-of-Words
- Πιθανοτικό Classification
  - Παράδειγμα Naive Bayes
  - Naive Bayes ως Language Model
- Classification Αξιολόγηση.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Resources

- Jurafsky, D. and H. Martin Justin, Chapter 4. "Naive Bayes and Sentiment Classification" Speech and Language Processing
- Jurafsky, D. and H. Martin Justin, Chapter 5. "Logistic Regression" Speech and Language Processing
- Python Tutorial on Text Classification using SKLearn: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423