ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Αλληλεπίδραση Ανθρώπου–Υπολογιστή

Β4. Επεξεργασία φυσικής γλώσσας με απλούς αλγορίθμους μηχανικής μάθησης

(2023-24)

Ίων Ανδρουτσόπουλος

http://www.aueb.gr/users/ion/

Τι θα ακούσετε σήμερα

- Μετατροπή κειμένων σε διανύσματα χαρακτηριστικών.
- Κατηγοριοποίηση κειμένων με απλούς αλγορίθμους επιβλεπόμενης μηχανικής μάθησης.
- Μέτρα αξιολόγησης αλγορίθμων κατηγοριοποίησης κειμένων.
- Ενθέσεις λέξεων (word embeddings).
- Συσταδοποίηση (clustering) κειμένων ή λέξεων.

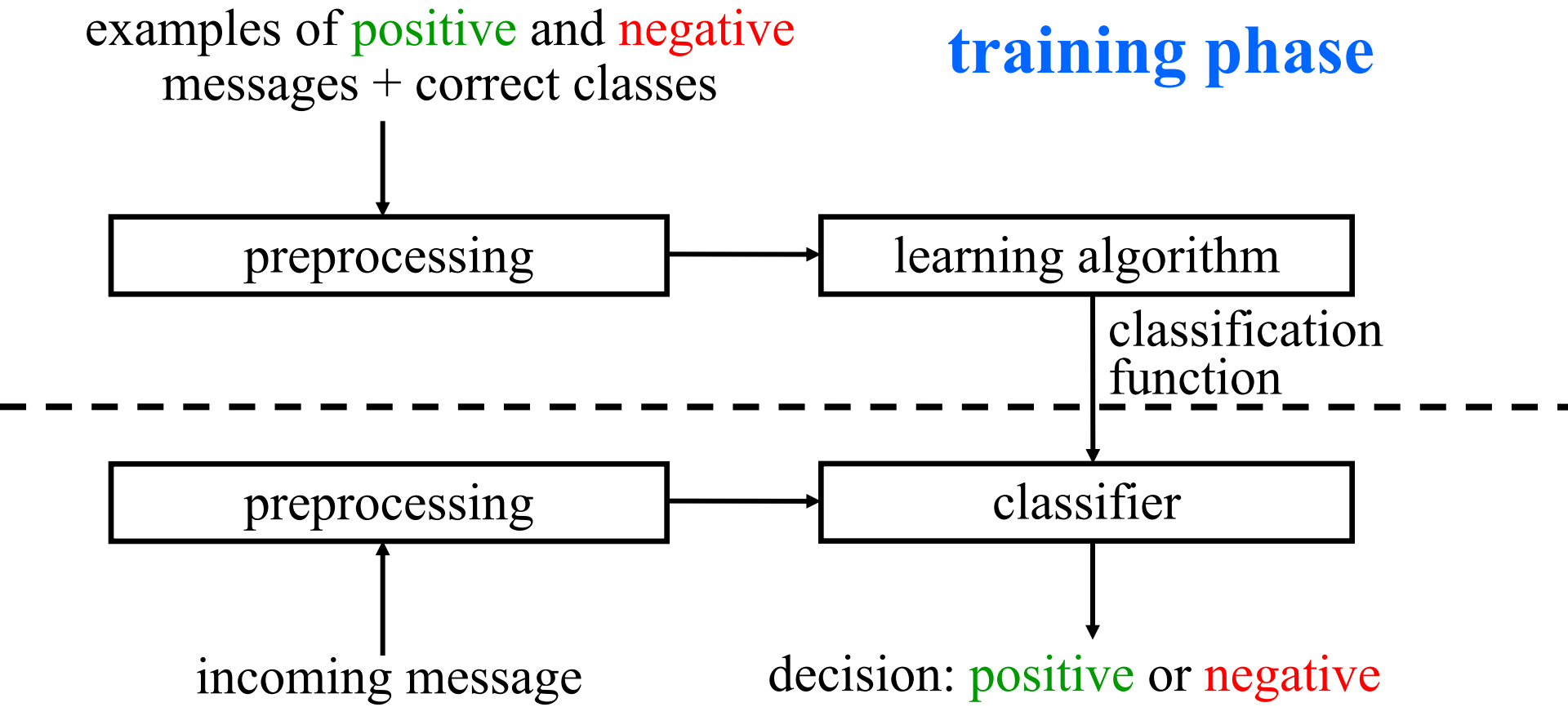
Example: spam filters

our highly successful multi – national company gives you an exclusive business that generates an extra weekly income of up to $ 600 or more … anyone can easily make money … if you wish to be removed from our list …

call for papers 9 th european workshop on natural language generation … is a subfield of natural language processing that generates texts in human languages from non-linguistic data or knowledge … for the systems to be successful …

- **Classifying messages** in **two classes.**
  - Spam ($C = 1$), ham ($C = 0$).
- More generally, ***n*** **classes**.
  - Financial news, politics, sports news (**possibly overlapping**).
  - Positive, negative, neutral **sentiment** (of tweets, reviews).

Spam filtering with supervised ML



use (or testing) phase

Text preprocessing

our highly successful multi – national company gives you an exclusive business that generates an extra weekly income of up to $ 600 or more … anyone can easily make money … if you wish to be removed from our list …

call for papers 9 th european workshop on natural language generation … is a subfield of natural language processing that generates texts in human languages from non-linguistic data or knowledge … for the systems to be successful …

< money:1, language:0, natural:0, $:1, adult:0, call:0, exclusive:1, **successful:1**, removed:1, **generates:1**, …>

< money:0, language:1, natural:1, $:0, adult:0, call:1, exclusive:0, **successful:1**, removed:0, **generates:1**, …>

- Alternatively the **features** may be **word (or *n*-gram) frequencies**, **TF-IDF** scores, **non-textual** information (e.g., attachments, colors).

Representing texts as bags of words

- **Boolean** vectors (contain 0, 1 values): **which words** of a vocabulary **occur or not** in the text.
- **Term frequency** (***TF***) vectors: **how frequent** each vocabulary word is in the text.
  - Possibly **divided by the number of tokens** of the text.
- ***TF-IDF*** vectors: **for each vocabulary word w_i**, the vector contains its $TF_i \cdot IDF_i$ **score**:
  - We want **frequent words of the text** that are **infrequent in the language** to have **large values** (they are important).

$$IDF_i = \log\left(\frac{Ndoc}{DF_i}\right)$$

Number of documents in a corpus.

Number of corpus documents containing w_i.

  - IDF_i (inverse document frequency) shows **how rare w_i is in the language**.

Επιλογή ιδιοτήτων

- Για **ποιες λέξεις** (ή φράσεις ή ...) θα έχουμε **ιδιότητες**;
  - 1ο βήμα: μόνο λέξεις που εμφανίζονται **τουλάχιστον** k φορές στα παραδείγματα εκπαίδευσης (π.χ. $k = 3$).
  - Συνήθως παραμένουν **χιλιάδες** λέξεις (ή φράσεις ή …).
  - Με χιλιάδες ιδιότητες: προβλήματα **ταχύτητας**, **ορθότητας** και **υπερεφαρμογής** με πολλούς αλγορίθμους μηχανικής μάθησης.
- Πόσο **αξίζει** κάθε υποψήφια ιδιότητα $\boldsymbol{X}$ (π.χ. X_{money});
  - $\boldsymbol{C = 1}$ (ανεπιθύμητο) ή $\boldsymbol{C = 0}$ (επιθυμητό).
  - Πόσο μειώνεται η **αβεβαιότητά** μας (εντροπία) για την τιμή της **τυχαίας μεταβλητής** $\boldsymbol{C}$, αν ξέρουμε την τιμή της X;
  - **Κέρδος Πληροφορίας** (Information Gain) IG(C, X). Βλ. προαιρετικά διαφάνειες μαθήματος «Τεχνητή Νοημοσύνη».
  - Υπάρχουν και άλλοι τρόποι **επιλογής/εξαγωγής ιδιοτήτων** (π.χ. χ^2, αναζήτηση βέλτιστου υποσυνόλου ιδιοτήτων, PCA/SVD).

Example of IG-selected Boolean features

Word of X_i	$P(X_i=1)$	$P(X_i=1\|C=0)$	$P(X_i=1\|C=1)$
!	0.484105	0.216129	0.828157
$	0.257947	0.040322	0.538302
language	0.247956	0.440322	0.002070
money	0.163487	0.001612	0.372670
remove	0.146230	0.001612	0.333333
free	0.309718	0.104838	0.573498
university	0.219800	0.374193	0.022774

Text preprocessing

our highly successful multi – national company gives you an exclusive business that generates an extra weekly income of up to $ 600 or more … anyone can easily make money … if you wish to be removed from our list …

call for papers 9 th european workshop on natural language generation … is a subfield of natural language processing that generates texts in human languages from non-linguistic data or knowledge … for the systems to be successful …

< money:1, language:0, natural:0, $:1, adult:0, call:0, exclusive:1, **successful:1**, removed:1, **generates:1**, …>

< money:0, language:1, natural:1, $:0, adult:0, call:1, exclusive:0, **successful:1**, removed:0, **generates:1**, …>

- Alternatively the **features** may be **word (or *n*-gram) frequencies**, **TF-IDF** scores, **non-textual** information (e.g., attachments, colors).

Αλγόριθμοι επιβλεπόμενης μάθησης

- **Διαδεδομένοι απλοί αλγόριθμοι επιβλεπόμενης μάθησης:**
  - Αφελείς ταξινομητές Bayes, k-NN, Logistic Regression, ID3, Random Forest, …
  - Βλ. προαιρετικά διαφάνειες μαθήματος «Τεχνητή Νοημοσύνη» (https://eclass.aueb.gr/courses/INF153/).
  - **Έτοιμες υλοποιήσεις**, π.χ. https://scikit-learn.org/.
  - Μπορούν να χρησιμοποιηθούν για **κατηγοριοποίηση κειμένων**, **αφού μετατρέψουμε τα κείμενα σε διανύσματα χαρακτηριστικών** (π.χ. χαρακτηριστικά συχνοτήτων λέξεων ή χαρακτηριστικά TF-IDF).

Δεδομένα εκπαίδευσης, ανάπτυξης, ελέγχου

- **Δεδομένα εκπαίδευσης** (training data):
  - Δεδομένα στα οποία εκπαιδεύεται ο αλγόριθμος μάθησης.
  - Συχνά δοκιμάζουμε να εκπαιδεύσουμε τον αλγόριθμο στο x% των δεδομένων εκπαίδευσης (x = 10%, ..., 100%).
- **Δεδομένα ανάπτυξης/ελέγχου** (development/test data):
  - Δεδομένα στα οποία ελέγχουμε την επίδοση της συνάρτησης που μάθαμε, κατά την ανάπτυξη/τελική δοκιμή.
  - **Διαφορετικά** από τα δεδομένα εκπαίδευσης, αλλά από τον ίδιο πληθυσμό (π.χ. κριτικές ίδιων ειδών προϊόντων).
- **Παραδείγματα μέτρων αξιολόγησης**:
  - **Ποσοστό ορθότητας** (accuracy): ποσοστό περιπτώσεων για τις οποίες η απόκριση του ταξινομητή είναι σωστή.
  - **Ποσοστό λάθους** (error rate): 1 – accuracy.
  - **Μέσο τετραγωνικό ή απόλυτο σφάλμα**, όταν προβλέπουμε τιμή (παλινδρόμηση) αντί για κατηγορία.

Evaluating classifiers

- **Accuracy** (correct decisions/total decisions) is **not always** a **good evaluation measure**!
  - If we have two classes and one is much more frequent (e.g., 80% of instances), a **majority classifier** that always classifies in the most frequent class will have an accuracy of 80%!
- **Precision of a class**:
  - **How many** of the instances **classified in the class** (true positives + false positives) **are true members** of the class (true positives).

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

- **Recall of a class**:
  - **How many** of the **true members** of a class (true positives + false negatives) are **classified in the class** (true positives).

Evaluating classifiers – continued

- **F-measure**: $$F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$
  - **Combination** of **precision** and **recall** (weighted harmonic mean).
  - For $\boldsymbol{\beta = 1}$, **equal importance** to **precision** and **recall**. (But the harmonic mean is closer to the min of the two values than the arithmetic mean.)
- **Averaging precision or recall** over ***n*** **classes**:
  - **Macro-averaging** (equal weight assigned to all classes):

$$\text{MacroPrecision} = \frac{1}{n}\sum_{i=1}^{n} \text{Precision}_i \qquad \text{MacroRecall} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i$$

  - **Micro-averaging** (frequent classes treated as more important):

$$\text{MicroPrecision} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FP_i} \qquad \text{MicroRecall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FN_i}$$

Precision-recall diagrams

- In many algorithms, we can **opt for higher precision** at the expense of **lower recall**, or **vice versa** by **tuning** a **threshold**.
  - In logistic regression: e.g., "spam" iff $P(C = 1|\vec{x}) > t$
  - For **different values** of the **threshold** t, we obtain **different pairs of precision-recall scores** (on test data).
  - The **larger** the **area under the curve** (**AUC** of Precision-Recall curve) the **better** the system.

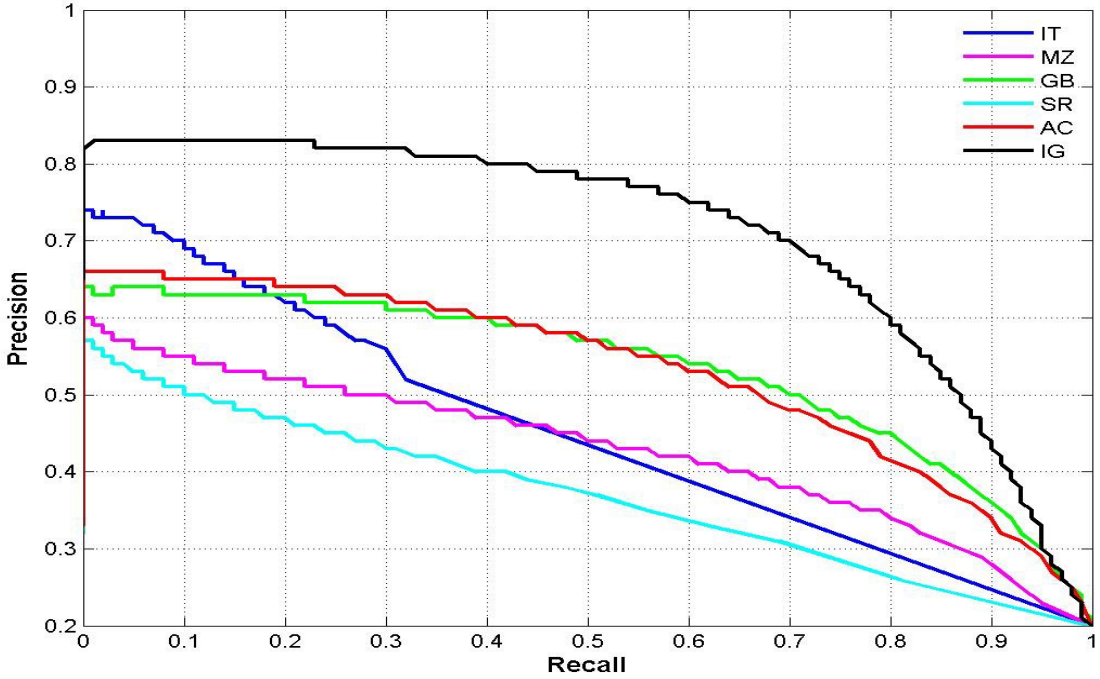


Image source: http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/Images/prcurve_1000.jpg

ROC curves

- Instead of Precision-Recall curves, it is also common to plot **Receiver Operating Characteristic** (ROC) curves.
  - **True Positive Rate** = $\frac{TP}{TP+FN}$ = **Sensitivity** = Recall of positive class
  - **False Positive Rate** = $\frac{FP}{FP+TN}$ = 1 – $\frac{TN}{TN+FP}$ = **1 – Specificity** = 1 – Recall of negative class
  - The **larger** the **AUC** (of ROC curve) the **better** the system.

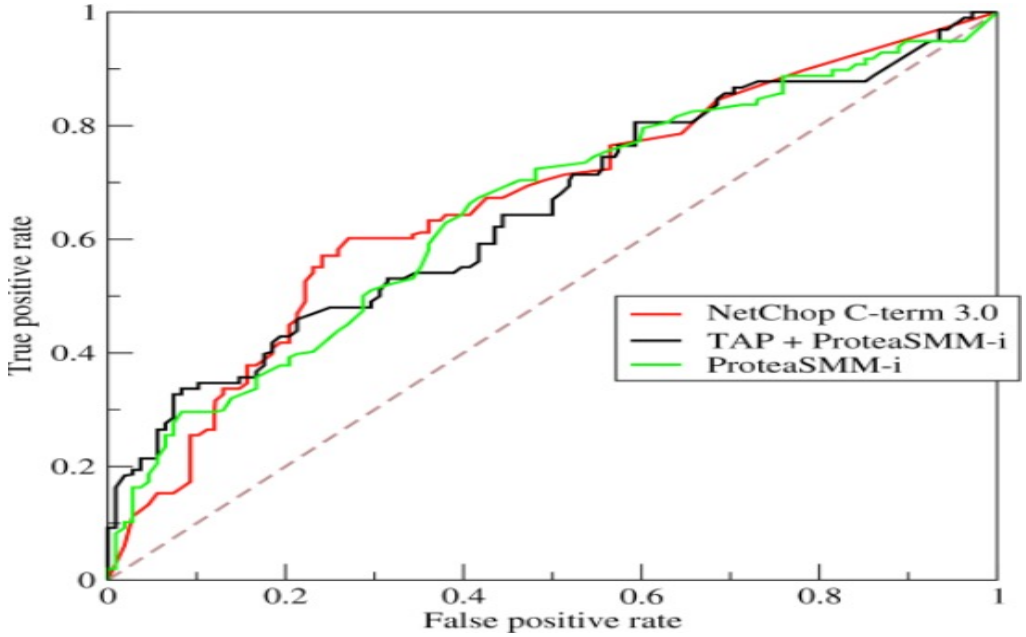


Image source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roccurves.png

Word embeddings

(produced by a method that produces **dense**, **sense-specific** word embeddings, then **projected** to **2 dimensions**)

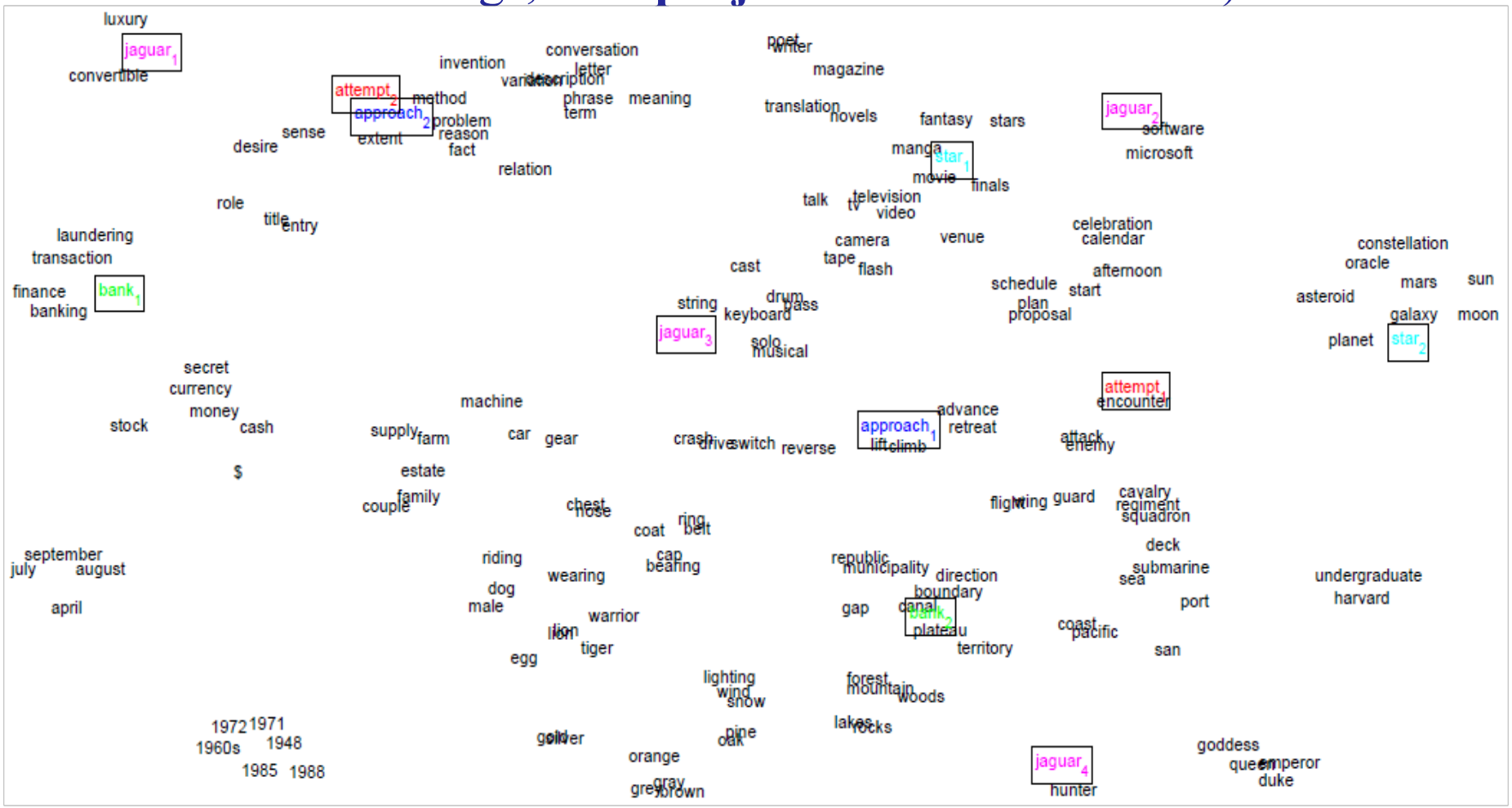


Image source: http://www.socher.org/uploads/Main/MultipleVectorWordEmbedding.png

Huan et al. 2012, “Improving Word Representations via Global Context and Multiple Word Prototypes”.

Embeddings of biomedical terms

Table 1 Closest words to the 30 most frequent words of the BioASQ question answering task, using the cosine similarity of the dense vectors to measure proximity. Relevant (closely related) words are shown in bold, possibly relevant in normal font, and irrelevant (or misspelled) words in strikeout.

protein	**proteins**	**a-anchoring**	**pka-anchoring**
thyroid	**thyroidal**	**nonthyroid**	hyperfunctioning
associated	**correlated**	**related**	**correlates**
hormone	**gh**	**luetinizing**	**fshluteinizing**
human	murine	mouse	~~immortalized~~
used	**utilized**	**employed**	**applied**
genes	**gene**	**paralogs**	**operons**
treatment	**therapy**	**treatments**	**treating**
disease	**diseases**	disease-like	~~mmrn1rs6532197~~
gene	**genes**	**pseudogene**	**gene-encoding**
heart	**cardiac**	**chf**	**congestive**
role	**roles**	~~plays~~	~~play~~
affect	alter	modify	impair
dna	**dnas**	bisulfite-treated	polymerase-mediated
histone	**histones**	**h4k16**	**h4**
involved	**implicated**	**participates**	**regulating**
list	**lists**	**listing**	to-do
proteins	**protein**	**polypeptides**	**hsp70s**
known	~~yet~~	~~presently~~	well-known
patients	**outpatients**	**subjects**	whom
present	this	~~aimed~~	~~our~~
cancer	**cancers**	**crc**	~~caner~~
receptor	**receptors**	**hmc5**	5-nonyloxytryptamine
regulate	**modulate**	**regulates**	**orchestrate**
cell	**cells**	**cancer-cell**	**sw1710**
coding	**5-noncoding**	**5-untranslated**	**3-noncoding**
inhibitors	**inhibitor**	**small-molecule**	**atp-competing**
many	**several**	**some**	**numerous**
related	**linked**	**associated**	**relate**
cardiomyopathy	**cardiomyopathies**	**myocardiopathy**	**dcm**

See http://bioasq.org/news/bioasq-releases-continuous-space-word-vectors-obtained-applying-word2vec-pubmed-abstracts

Word embeddings of business terms

(produced with **word2vec**, then **projected to 2D** using **UMAP**)

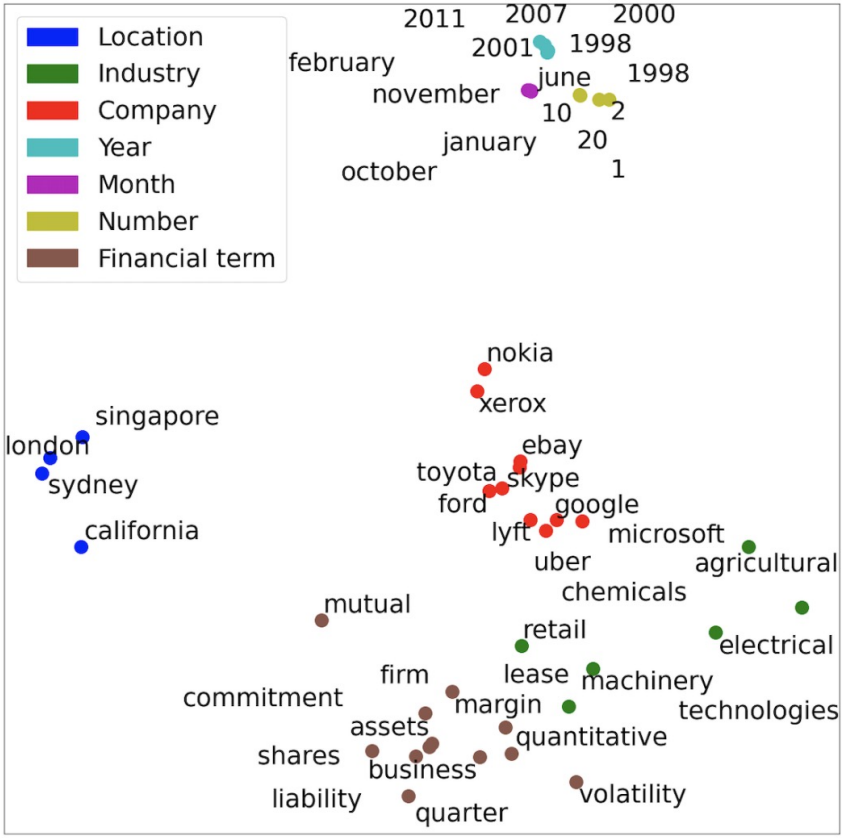


Image from Lukas et al., "EDGAR-CORPUS: Billions of Tokens Make The World Go Round", EcoNLP workshop, EMLP 2021. https://arxiv.org/abs/2109.14394

For a quick intro to **UMAP** (and **t-SNE**) check: https://www.youtube.com/watch?v=6BPl81wGGP8

Ενθέσεις λέξεων (word embeddings)

- **Ενθέσεις λέξεων**:
  - **Σχετικές λέξεις** απεικονίζονται σε **κοντινά διανύσματα**.
  - Τα διανύσματα συνήθως είναι **πυκνά** (ελάχιστα μηδενικά), με **100-300 διαστάσεις**.
  - **Ενώ σε 1-hot διανύσματα λέξεων**, έχουμε τόσες **διαστάσεις όσο το μέγεθος του λεξιλογίου** και **μόνο μία συνιστώσα** (αυτή που αντιστοιχεί στη συγκεκριμένη λέξη) είναι **μη μηδενική**.
- Κατασκευή **ενθέσεων λέξεων**.
  - Υπάρχουν **εργαλεία** που παράγουν ενθέσεις λέξεων από **μεγάλα σώματα κειμένων** (π.χ. Wikipedia).
  - Π.χ. **Word2vec** (https://code.google.com/archive/p/word2vec/), **GloVe** (https://nlp.stanford.edu/projects/glove/).
  - Εναλλακτικά **μπορούμε να μάθουμε** (ή να τροποποιήσουμε) ενθέσεις λέξεων **με δικά μας νευρωνικά δίκτυα**.

Κεντροειδή ενθέσεων λέξεων

- Μπορούμε να παραστήσουμε **κάθε κείμενο *T*** (ακολουθία λέξεων) $w_1, \dots w_d$ ως το **κεντροειδές** των **ενθέσεων λέξεων** του κειμένου:

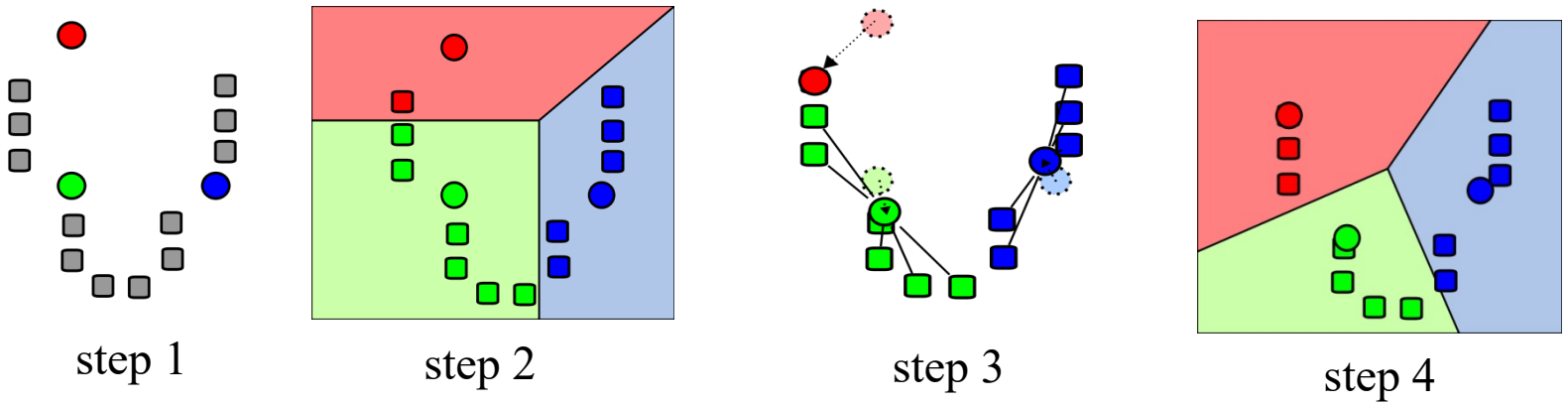
$$\vec{T} = \frac{1}{d}\sum_{i=1}^{d} \vec{w}_i = \frac{\sum_{j=1}^{|V|} \vec{w}_j \cdot TF(w_j, T)}{\sum_{j=1}^{|V|} TF(w_j, T)}$$

- Ή (καλύτερα) λαμβάνοντας υπόψη τις **τιμές *IDF*** των λέξεων:

$$\vec{T} = \frac{\sum_{j=1}^{|V|} \vec{w}_j \cdot TF(w_j, T) \cdot IDF(w_j)}{\sum_{j=1}^{|V|} TF(w_j, T) \cdot IDF(w_j)}$$

- Μπορούμε να **κατατάξουμε κείμενα σε κατηγορίες** κατατάσσοντας τα **κεντροειδή των κειμένων**.
- Θα δούμε καλύτερους τρόπους αργότερα…

Clustering with *k*-means



- Start with k **random centroids** (one per desired cluster).
- Place **each instance** into the **cluster** with the **closest centroid**.
- **Re-compute** the **centroids**. **Repeat** until convergence.

➢ **Unsupervised learning**. Can be made **semi-supervised** (how?).

➢ Produces **hard clusters**, unlike EM's $P(C = i|\vec{X})$.

➢ Tries to minimize the **sum of the distances** of the **instances** to the **centroids** of their clusters.

➢ May find a **local minimum**. Sensitive to the initial random centroids. **Restart** several times with **different initial random centroids**.

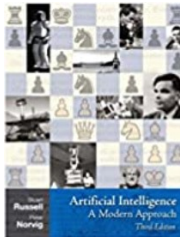
Images from: http://en.wikipedia.org/wiki/K-means_clustering

Clustering documents and/or words

- We can **cluster documents** (e.g., their *TF-IDF* vectors).
  - For example, hoping to get a **view** of their **topics**.
  - More elaborate **topic modeling** methods (e.g., **LDA**) exist.
- We can **cluster word embeddings**.
  - For example, to **replace words** by their **clusters** in **BOW** representations of documents (fewer features).

Βιβλιογραφία

- Για περισσότερες πληροφορίες, δείτε τις διαφάνειες των διαλέξεων 15 – 18 του μαθήματος «Τεχνητή Νοημοσύνη».
  - https://eclass.aueb.gr/courses/INF153
- Αν έχετε από το μάθημα της ΤΝ το βιβλίο των Russel & Norvig «Τεχνητή Νοημοσύνη – Μια σύγχρονη προσέγγιση», 4η έκδοση, Κλειδάριθμος, 2021, μπορείτε να συμβουλευτείτε το κεφάλαιο 19.



- Μπορείτε να συμβουλευτείτε και την 3η έκδοση του βιβλίου «Speech and Language Processing» των Jurafsky & Martin (υπό προετοιμασία), που διατίθεται δωρεάν.
  - http://web.stanford.edu/~jurafsky/slp3/