Clause Classification

Yogesh Kulkarni

Icertis

Agenda

- Introduction
  - Text Classification
  - Clause Classification
- Dataset
- Vectorization
- Classification
- Metrics

Introduction to Text Classification

Is this spam?

From: U.S. Bank <service@usbank.com>
Subject: **Customer Service**
Date: December 8, 2008 5:25:15 AM PST
To: undisclosed-recipients:;

This is a reminder that your U.S. Bank Account needs to be verified.
To continue using your card, please verify your account immediately.

To verify your account, please click the link below, log in and follow the provided steps:

http://www4.usbankv.com/internetBanking/?LoginRouter

Regards,
U.S. Bank

(Ref: Text Classification - The Naïve Bayes algorithm)

Positive or negative movie review?

unbelievably disappointing

Full of zany characters and richly applied satire, and some great plot twists

this is the greatest screwball comedy ever filmed

It was pathetic. The worst part about it was the boxing scenes.

(Ref: Text Classification - The Naïve Bayes algorithm)

What is the subject of this article?

MEDLINE Article

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

(Ref: Text Classification - The Naïve Bayes algorithm)

Introduction to Clause Classification

What is a Contract?

- A legally binding agreement between two or more persons which the courts will enforce
- Generates rights and obligations that may be enforced by courts
- Agreement arises as a result of offer and acceptance

Requirements for a valid contract

- Parties must have legal capacity to enter contracts
- One party must make a binding offer to the other, and the offer must be accepted by the other party
- Consideration
- Agreement must be genuine
- In some cases, the contract must be made in a particular form
- The object of the contract must not be disapproved by the law

- In summary, need:
  - Offer
  - Acceptance
  - Consideration - An exchange of promises to perform acts in the future

Contents of the contract

- Terms - Conditions or duties which must be carried out as part of a contract, arrangemens which must be made before a contract is valid
  - Express terms: Terms that the parties have specifically agreed to
  - Implied terms: e.g., goods are of satisfactory quality and are fit for their purpose
  - Conditions: If a party does not carry it out, you not only have the right to claim damages, but also to treat the contract as terminated
  - Warranties: If you have not carried out your obligations under a warranty, the other party has the right to sue you for damages, but not to terminate the contract
  - Exclusion terms
- Termination:
  - Performance
  - Express agreement
  - Breach
- Text/Language used to describe the Term is called “Clauses”.

What are clauses?

- All the provisions for a contract are detailed in clauses: who gets paid, who does the work, and what happens if one party backs out of the contract.
- Clauses are specific provisions or sections in your contract that address a specific aspect of the agreement.
- Clauses clearly define each party’s duties, rights, and privileges under the terms of the contract.

Example Clauses: Exemptions

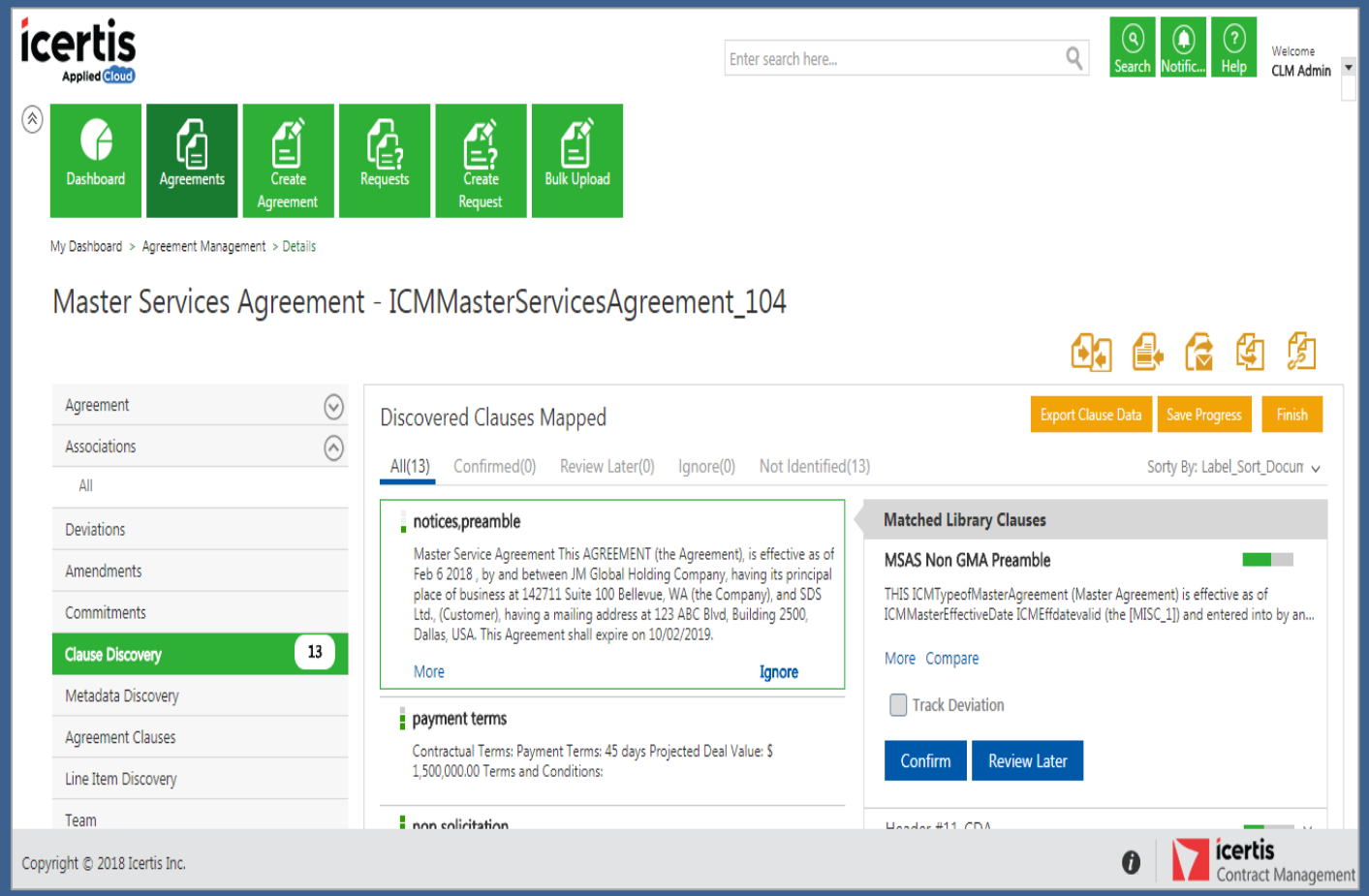
- Exemption clause is a provision in a contract that limits the liability of one party.
- Exemption types:
  - Exclusion: stipulates that one of the parties will not be liable to the other in certain specified situations or circumstances. For example, an exclusion clause in a life insurance contract may state that a death by suicide is excluded.
  - Indemnity: one party agrees to indemnify, or answer for, liability or losses incurred by another party.
  - Force Majeure (French: “Superior Forces”): Exempts the parties from liability if they are prevented from performing their contractual obligations by unforeseeable circumstances beyond their control.

Example Clauses

- Escalation clause is a provision in a contract that allows for one party to increase the contracted-for prices or wages under certain specified conditions.
- Penalty clause in a contract is a provision that requires a party that has not fulfilled its contractual obligations to pay specified compensation to the other party for the breach.
- Confidentiality clauses bar the receiving party from disclosing the information specified except under certain specified circumstances.

ICM DiscoverAI App

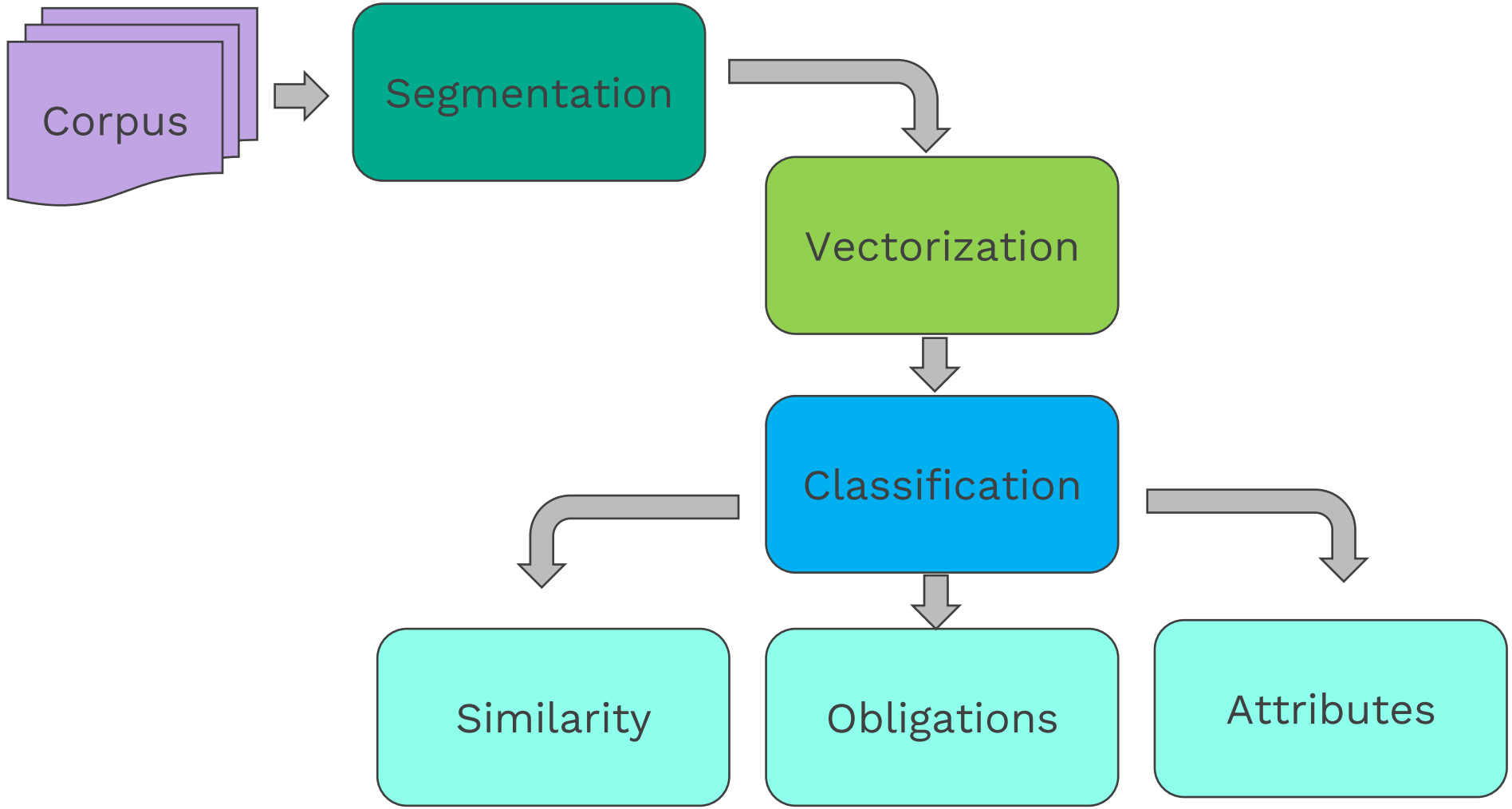
AI Powered NLP & Semantics Driven Clause Discovery, Deviations and Verifications



Powerful capabilities:

- Image/PDF Conversion
- Clause and Attribute Discovery
- Accurate Clause Matching
- Deviation Tracking
- Attribute Verification
- Driven by NLP & Machine Learning

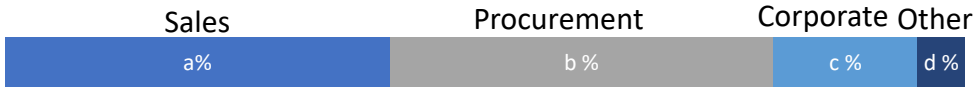
Clause Discovery Flow and Applications



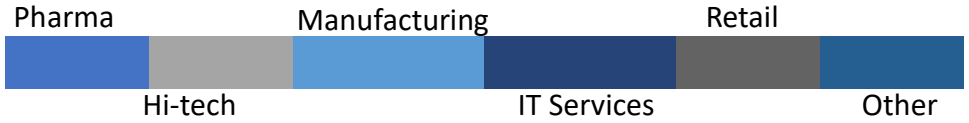
Dataset

Corpus Diversity

Distribution by function

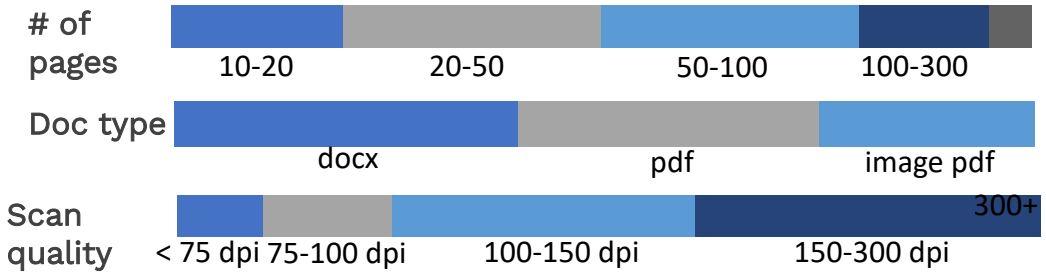


Distribution by Sector



Distribution by Region

Distribution by Complexity



Vectorization

Why Vectorize?

- Text processing with current NNs requires encoding into vectors.
- Prior work:
  - Learning representations by back-propagating errors. (Rumelhart et al., 1986)
  - A neural probabilistic language model (Bengio et al., 2003)
  - NLP (almost) from Scratch (Collobert & Weston, 2008)
  - A recent, even simpler and faster model: word2vec (Mikolov et al. 2013)

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Approaches

- One-hot encoding (Word):
  - N words encoded by length N vectors.
  - A word gets a vector with exactly one entry = 1, others 0.
  - Very space inefficient, no natural language structure.
- Bag of words (doc):
  - Collection of words (along with number of occurrences).
  - No word order, no natural language structure.
- Tf-Idf (doc)
  - Less sparse
  - No word order but unique words have more weights.
- Word Embeddings (word, doc):
  - Predict words using context
  - Considers order

Word embeddings: properties

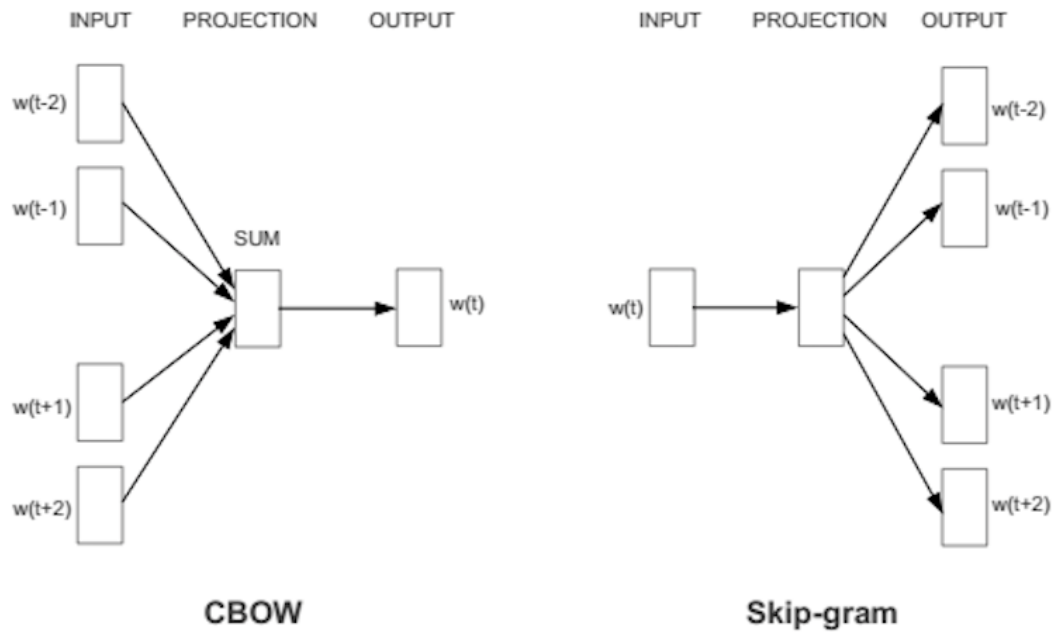
- Need to have a function W(word) that returns a vector encoding that word.

- Similarity of words corresponds to nearby vectors.
  - Director – chairman, scratched – scraped

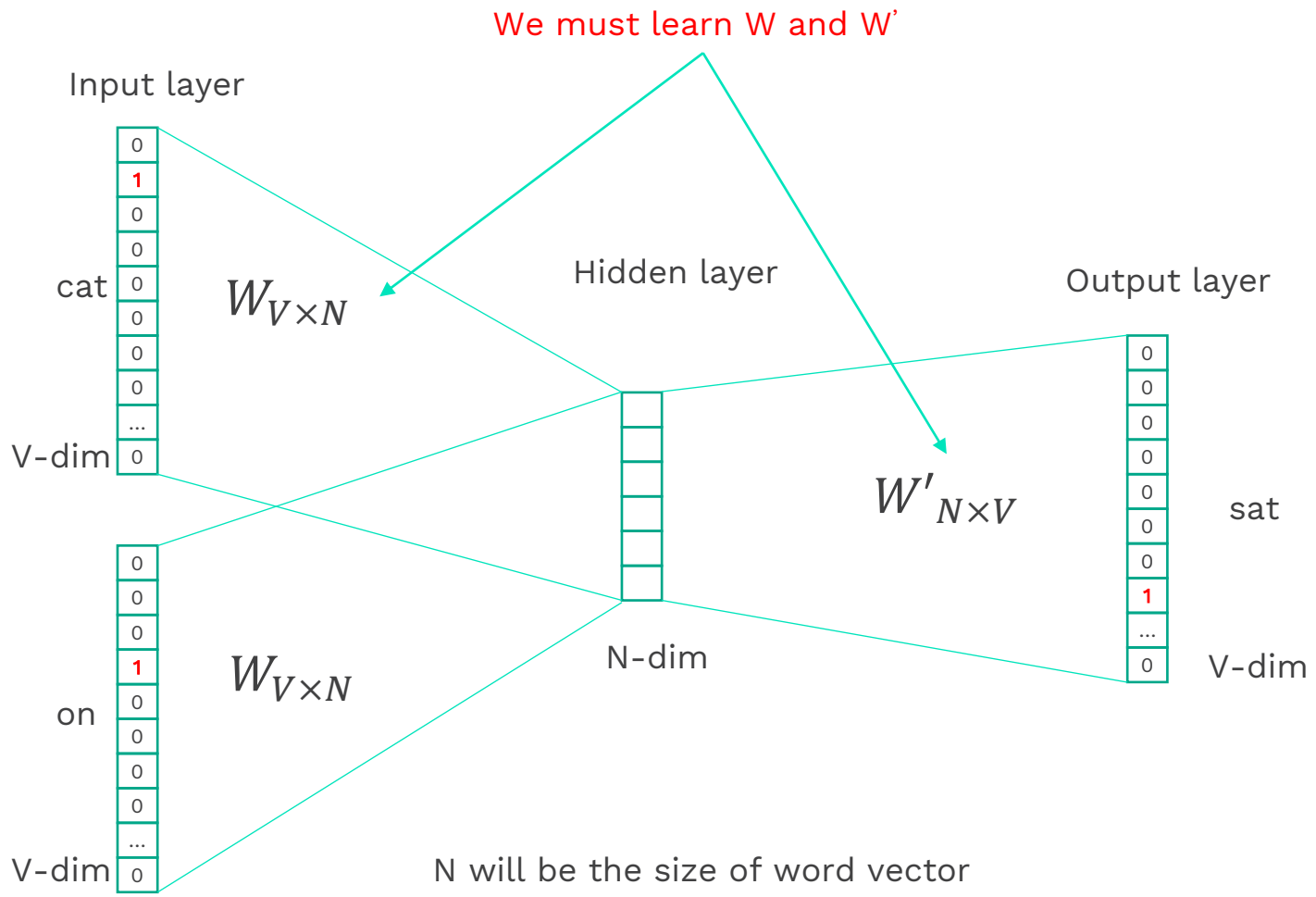
- Relationships between words correspond to difference between vectors.
  - Big – bigger, small – smaller

word2vec

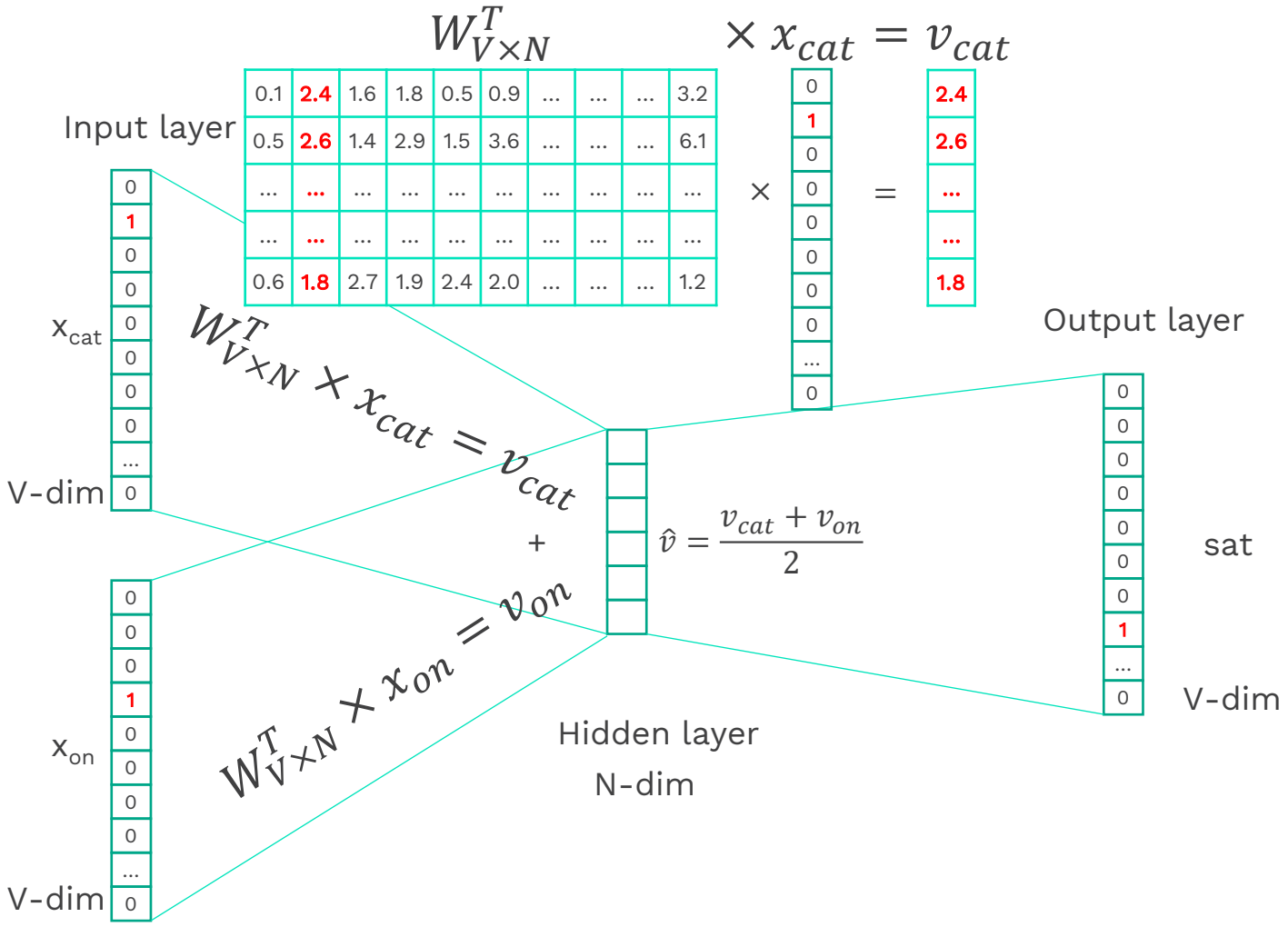
- Predict words using context
- Two versions: CBOW (continuous bag of words) and Skip-gram

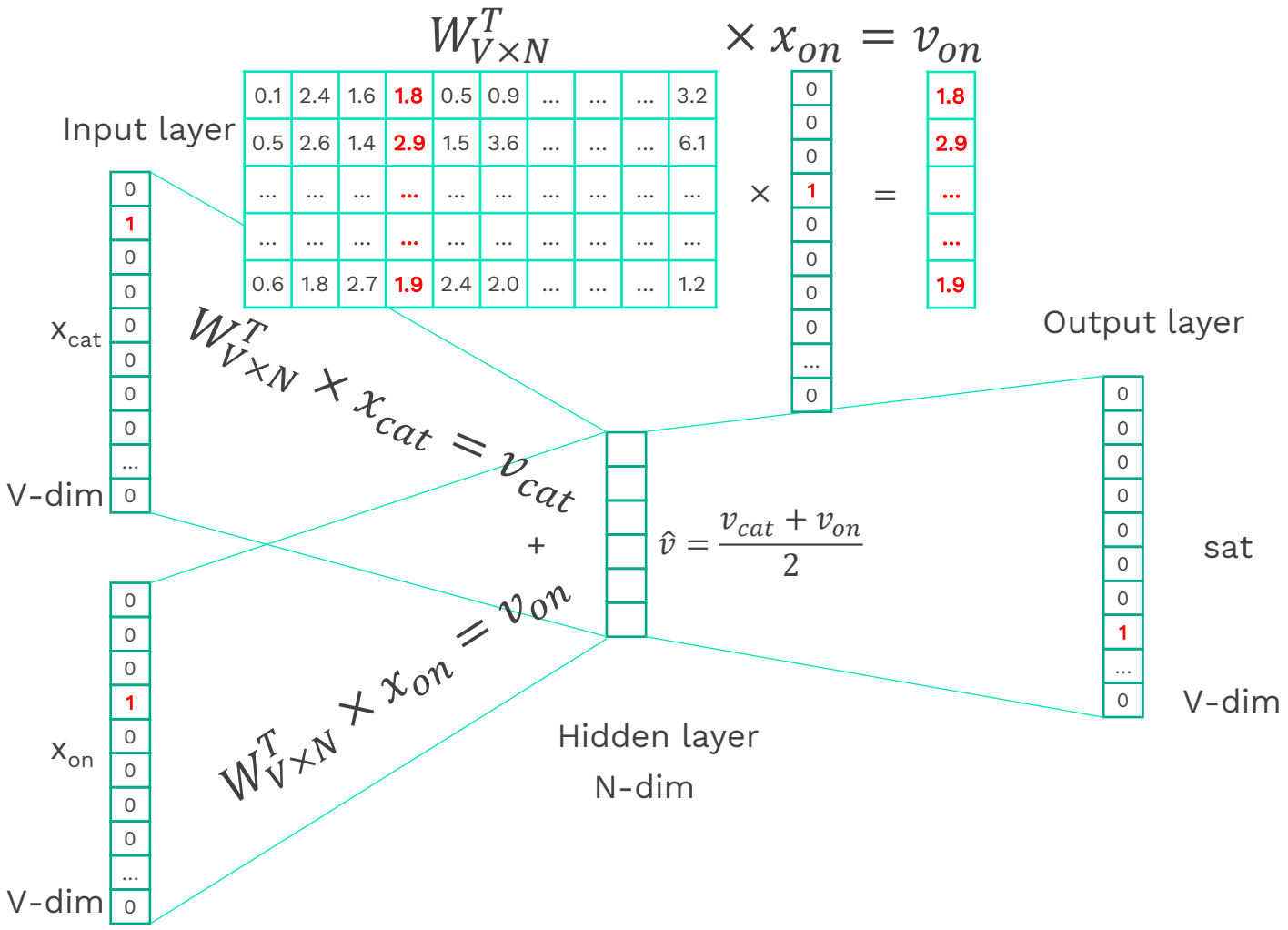


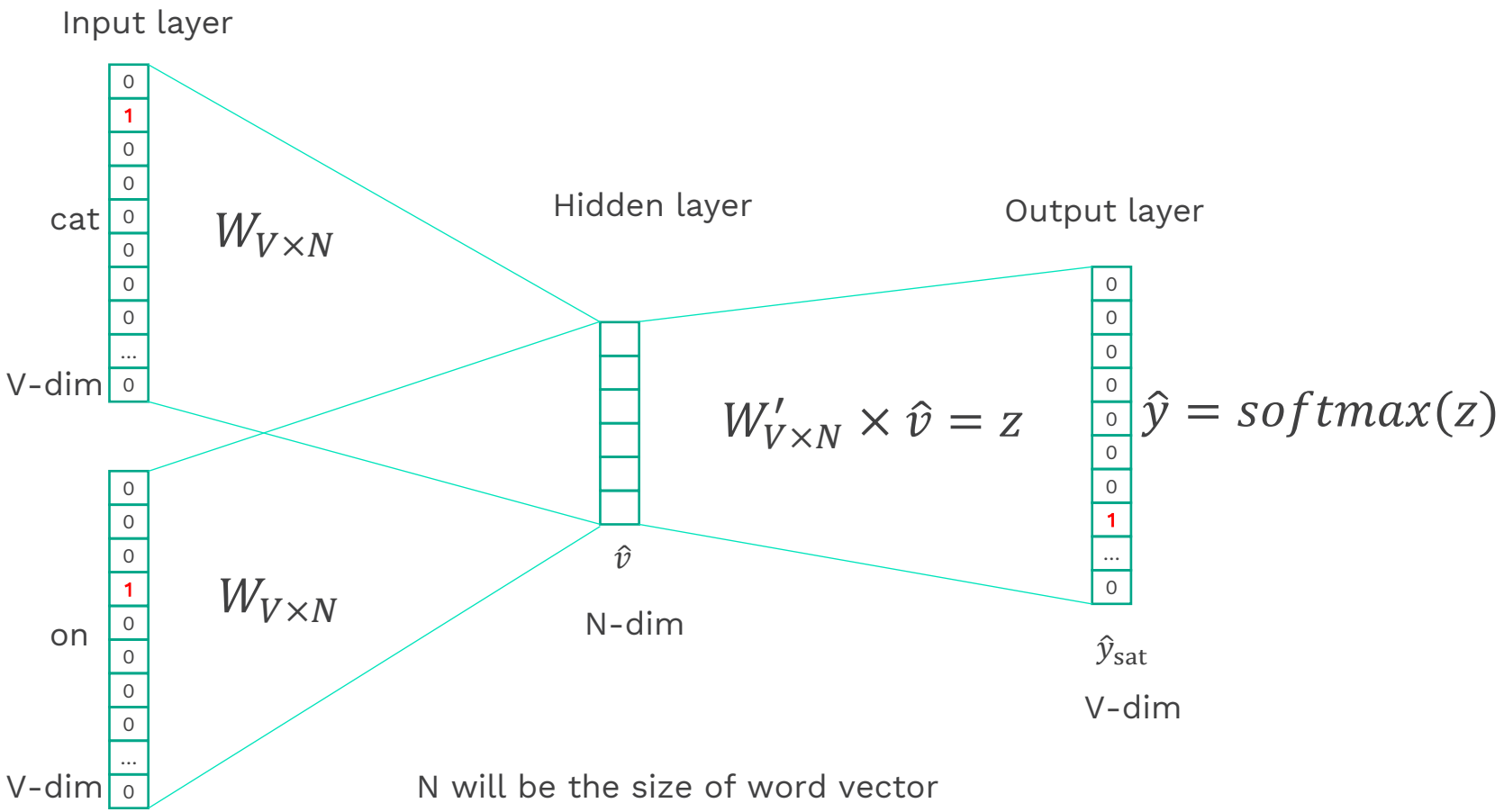
https://skymind.ai/wiki/word2vec



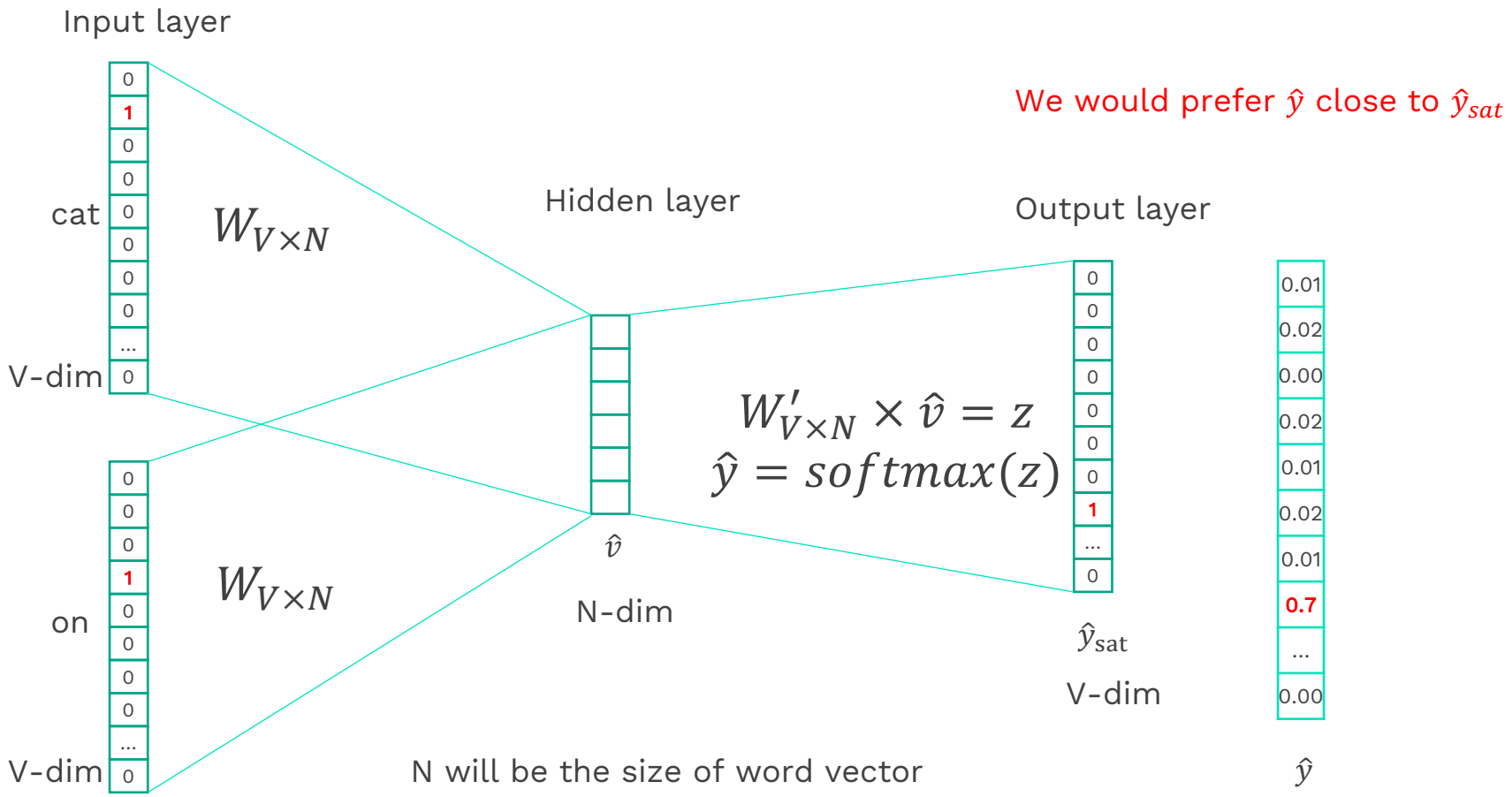
www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx


$W^T_{V\times N}$
$\times x_{cat} = v_{cat}$
Input layer
0.1 2.4 1.6 1.8 0.5 0.9 ... ... 3.2
0.5 2.6 1.4 2.9 1.5 3.6 ... ... 6.1
... ...
... ...
0.6 1.8 2.7 1.9 2.4 2.0 ... ... 1.2
0 1 0 0 0 0 0 0 0 0
×
=
2.4 2.6 ... ... 1.8
Output layer
x_{cat}
V-dim
$W^T_{V\times N} \times x_{cat} = v_{cat}$
+
$W^T_{V\times N} \times x_{on} = v_{on}$
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
sat
x_{on}
V-dim
Hidden layer
N-dim


$W^T_{V\times N}$
$\times x_{on} = v_{on}$
Input layer
0.1 2.4 1.6 1.8 0.5 0.9 ... ... 3.2
0.5 2.6 1.4 2.9 1.5 3.6 ... ... 6.1
... ...
... ...
0.6 1.8 2.7 1.9 2.4 2.0 ... ... 1.2
0 0 0 1 0 0 0 0 ... 0
1.8 2.9 ... ... 1.9
$W^T_{V\times N} \times x_{cat} = v_{cat}$
x_{cat}
V-dim
+
$W^T_{V\times N} \times x_{on} = v_{on}$
x_{on}
V-dim
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
Hidden layer
N-dim
Output layer
0 0 0 0 0 0 0 1 ... 0
sat
V-dim



www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx



www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx

Some interesting results

Word Analogies

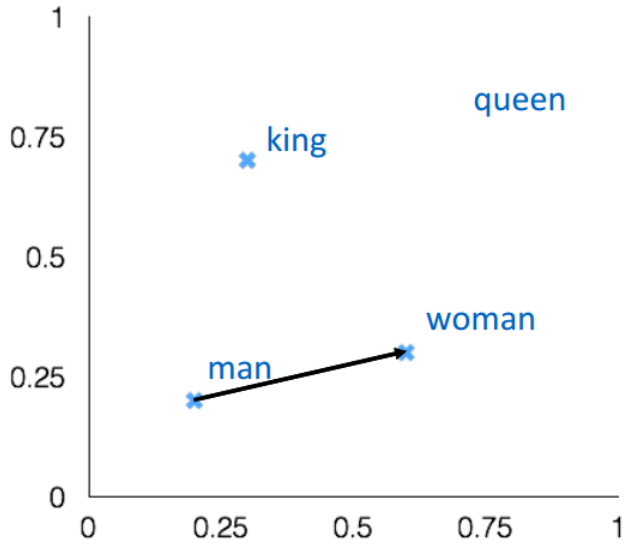
Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:?

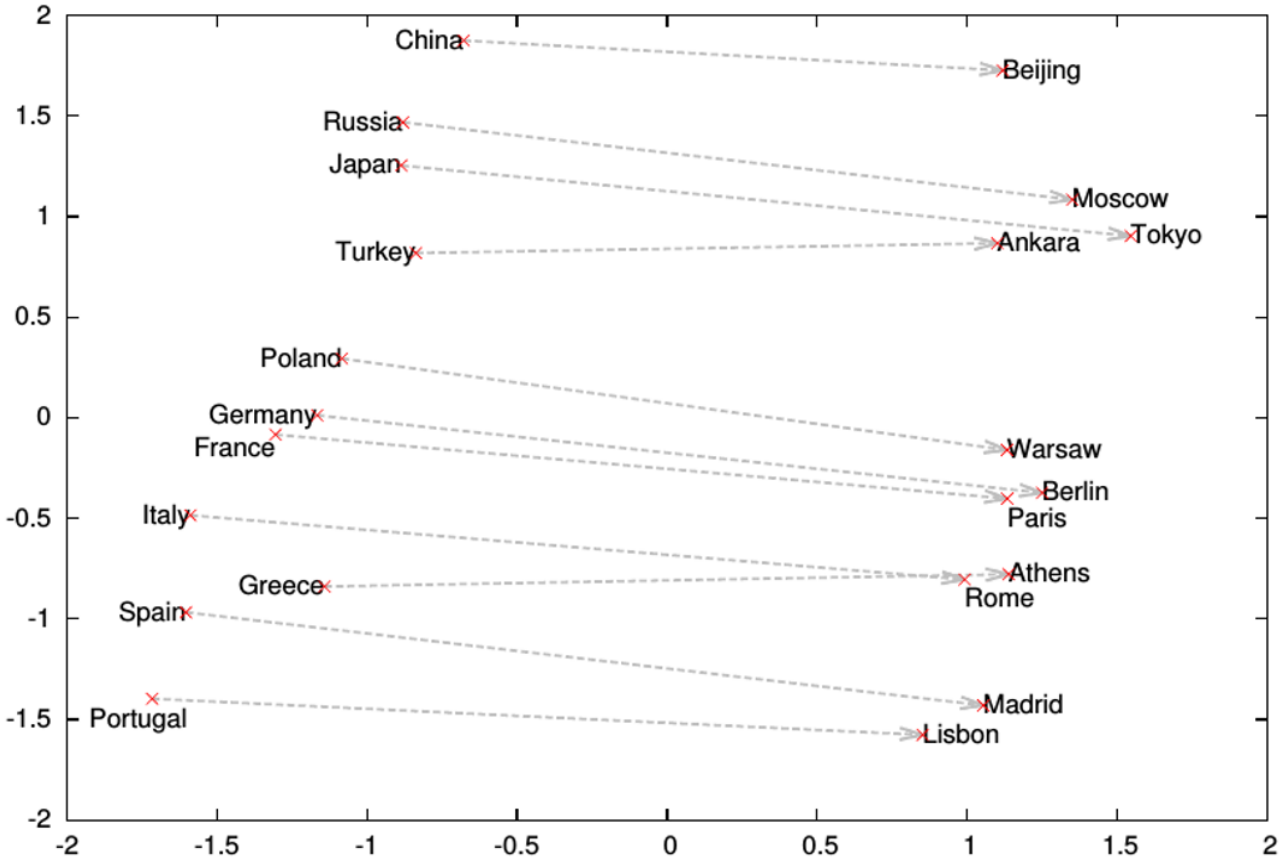
$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Word analogies



www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx

Skip gram/CBOW intuition

- Similar “contexts” (that is, what words are likely to appear around them), lead to similar embeddings for two words.
- One way for the network to output similar context predictions for these two words is if *the word vectors are similar*. So, if two words have similar contexts, then the network is motivated to learn similar word vectors for these two words!

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word2vec shortcomings

- **Problem:** 10,000 words and 300 dim embedding gives a large parameter space to learn. And 10K words is minimal for real applications.
- Slow to train, and need lots of data, particularly to learn uncommon words.

Classification

Classification methods: 1. Manual

- Manual classification was used by Yahoo in the beginning of the web. Also: ODP, PubMed
- Very accurate if job is done by experts
- Consistent when the problem size and team is small
- Scaling manual classification is difficult and expensive.
- → We need automatic methods for classification.

Classification methods: 2. Rule-based

- Our Google Alerts example was rule-based classification.
- There are IDE-type development environments for writing very complex rules efficiently. (e.g., Verity)
- Often: Boolean combinations (as in Google Alerts)
- Accuracy is very high if a rule has been carefully refined over time by a subject expert.
- Building and maintaining rule-based classification systems is cumbersome and expensive.

Classification methods: 3. Supervised Machine Learning

- Input:
  - a document d $\in$ D
  - a fixed set of classes C = {c1, c2,…, cK}
  - A training set of N hand-labeled documents T = { (d1,c1),….,(dN,cN) }
- Output:
  - a learned classifier γ: D → C

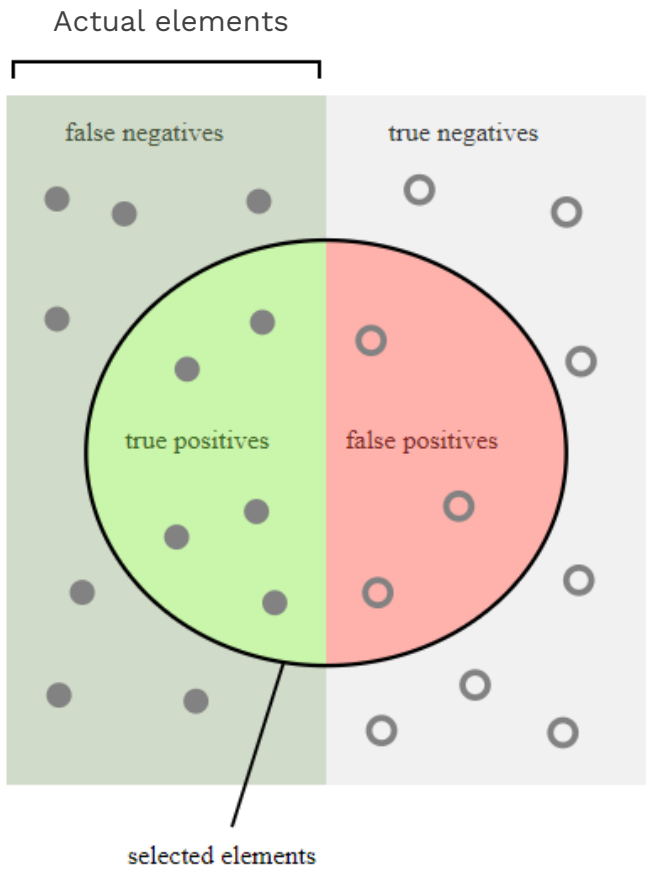
- Naive Bayes, kNN, SVMs
- No free lunch: requires hand-classified training data

Metrics

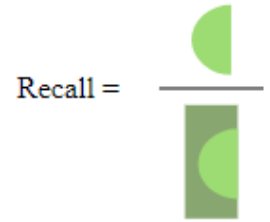
Evaluating classification

- Evaluation must be done on test data that are independent of the training data (usually a disjoint set of instances).
- It's easy to get good performance on a test set that was available to the learner during training (e.g., just memorize the test set).
- Measures: Precision, recall, F1, classification accuracy

Recap: Accuracy Measurements

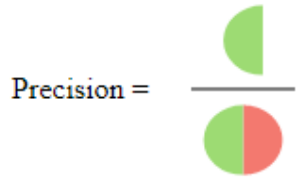


Recall – Measure of coverage
How many actual elements are selected



$$\text{Recall} = \frac{tp}{tp + fn}$$

Precision – Measure of correctness
How many selected elements are correct



$$\text{Precision} = \frac{tp}{tp + fp}$$

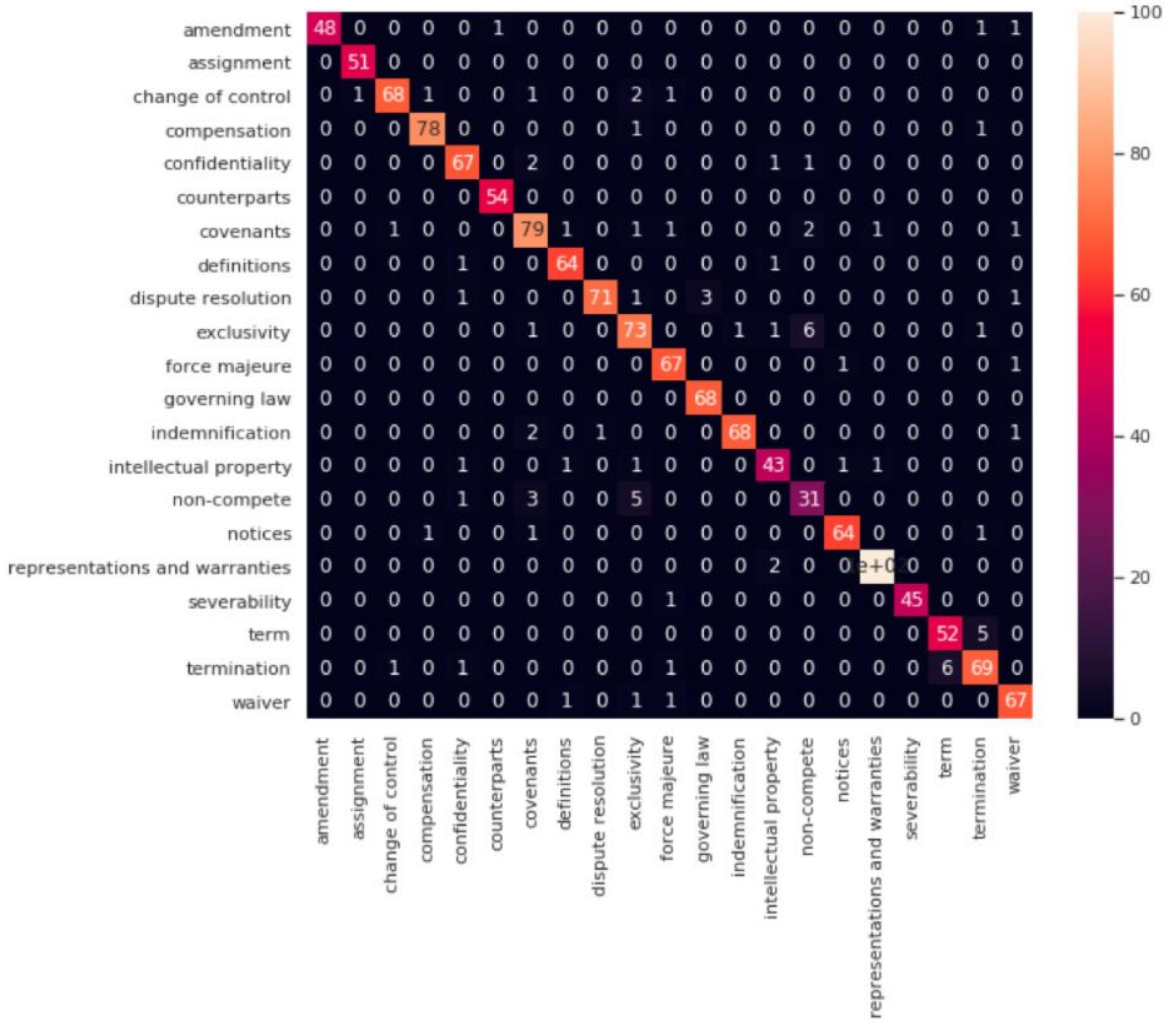
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Accuracy – Most intuitive parameter of effectiveness but fails when dataset is not balanced

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1 score – the balancing act
Measure to check effective of model without being biased towards precision or recall

Confusion Matrix



Next steps

- ✓ You can get soft copies of Reference materials/this ppt
- ✓ Apply at least a few concepts/Techniques

Questions?

Thank You

Icertis