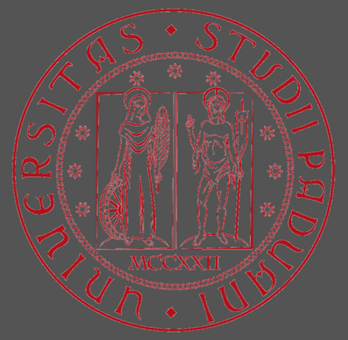
Master's Thesis in Data Science

Academic Year 2023-2024



Word2Box: Analysis And Exploration Of A Geometric Word Embedding Algorithm

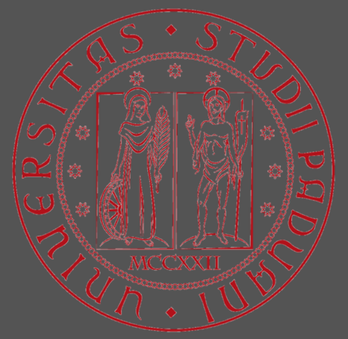
Master Candidate: Chiara Bigarella

Thesis Supervisor: Giorgio Satta

Department of Mathematics

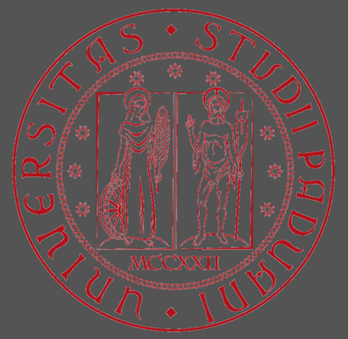
Università degli Studi di Padova

OUTLINE



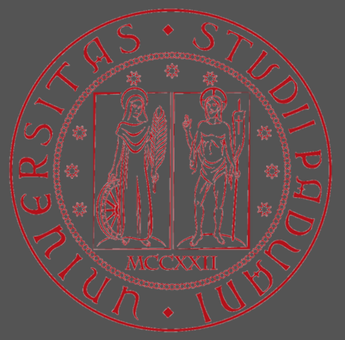
1. Word Embeddings
2. Geometric Representations
    a. Probabilistic Box Embeddings
3. Word2Box
4. Experiments
    a. Dataset
    b. Bugfixing
    c. Documentation
5. Future work

WORD EMBEDDINGS



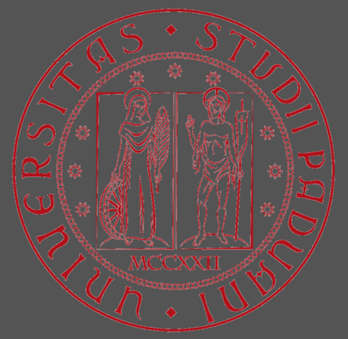
- represent words in a mathematical form, making it possible for NLP algorithms to process them
- map discrete words to numerical vectors in a continuous vector space
- based on the **distributional hypothesis** ➡ words that occur in similar contexts tend to have similar meanings

WORD EMBEDDINGS



Static Word Embeddings VS **Dynamic Word Embeddings**

WORD EMBEDDINGS



Static Word Embeddings VS **Dynamic Word Embeddings**

- univocally associate a fixed, precomputed embedding vector to each *word type* in the vocabulary *V*
- easy to implement, train and use ✅
- not able to represent *polysemic words* ❌
- not useful for *word sense disambiguation* ❌

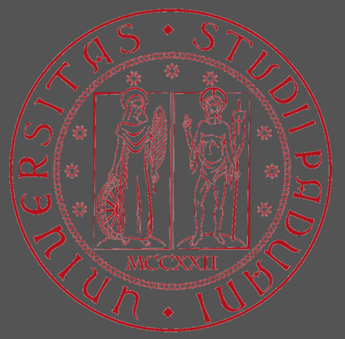
WORD EMBEDDINGS

Static Word Embeddings ⟷ VS ⟷ Dynamic Word Embeddings

- univocally associate a fixed, precomputed embedding vector to each *word type* in the vocabulary *V*
- easy to implement, train and use ✓
- not able to represent *polysemic words* ✗
- not useful for *word sense disambiguation* ✗

- the same *word type* has different embeddings, depending on the **context** of its occurrences
- dynamically computed by pre-trained LLM, finetuned on the downstream NLP task at hand
- able to represent *polysemic words* ✓
- computationally expensive ✗

WORD EMBEDDINGS



Static Word Embeddings

VS

Dynamic Word Embeddings

- univocally associate a fixed, precomputed embedding vector to each *word type* in the vocabulary *V*
- easy to implement, train and use
- not able to represent *polysemic words*
- not useful for *word sense disambiguation*

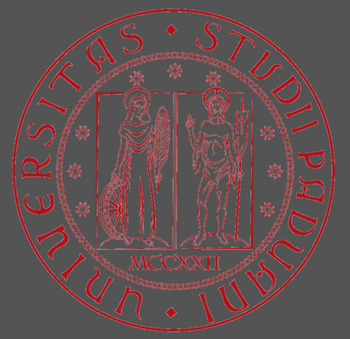
- the same *word type* has different embeddings, depending on the **context** of its occurrences
- dynamically computed by pre-trained LLM, finetuned on the downstream NLP task at hand
- able to represent *polysemic words*
- computationally expensive

GEOMETRIC REPRESENTATIONS

Limitations of Vector Embeddings

- do not express **uncertainty** about the target concept ⊗
- do not model **asymmetric relations** because they use **symmetric distance functions** such as dot product, cosine similarity, or Euclidean distance ⊗

GEOMETRIC REPRESENTATIONS



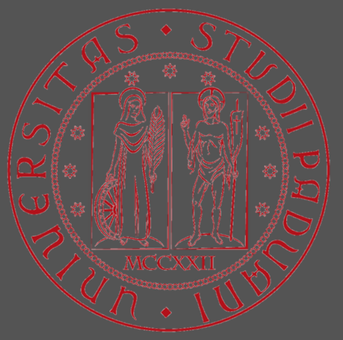
Limitations of Vector Embeddings

- do not express **uncertainty** about the target concept ✗
- do not model **asymmetric relations** because they use **symmetric distance functions** such as dot product, cosine similarity, or Euclidean distance ✗

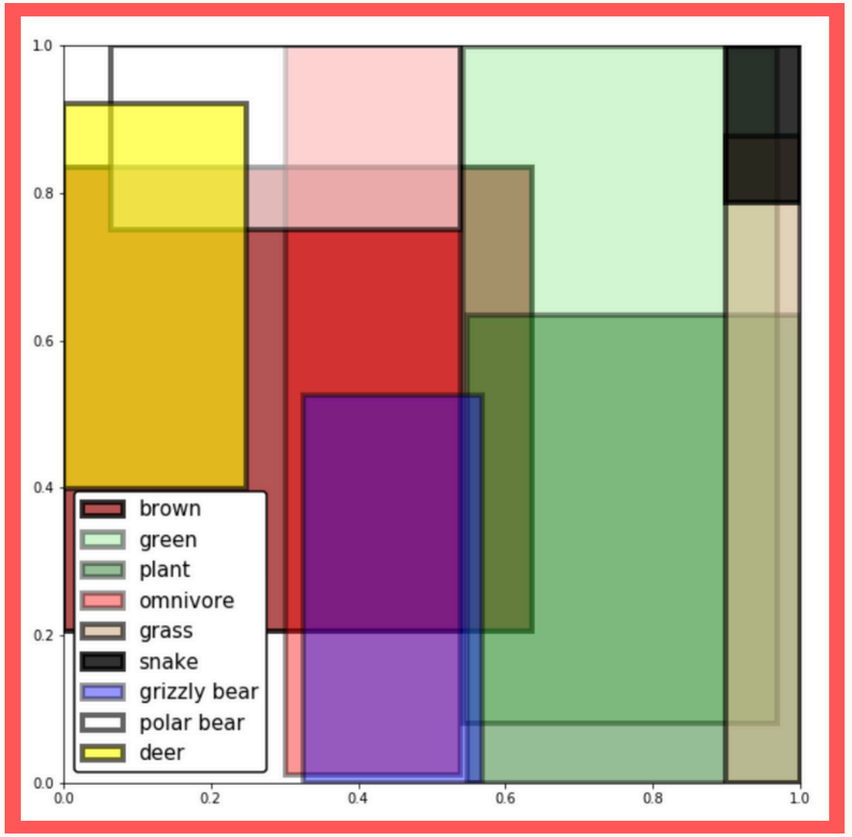
Geometric Representations

- represent entities as geometric objects in a high-dimensional space
- suitable to express **relationships** in the domain ✓
- able to represent **polysemy** ✓
- able to represent **asymmetry** ✓
- able to answer complex queries ✓

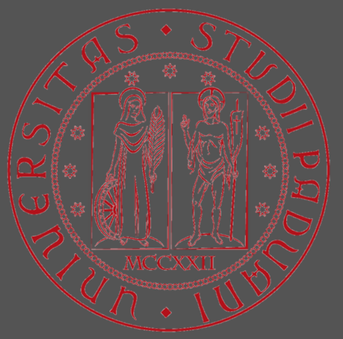
Probabilistic Box Embeddings



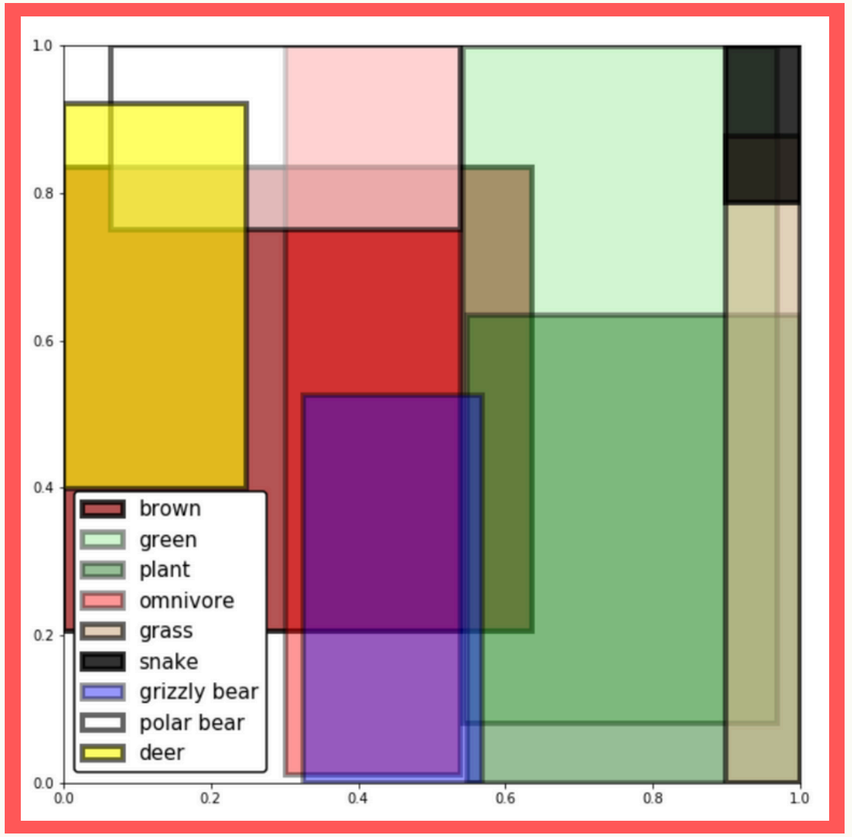
- A **geometric embedding** algorithm proposed in 2018 by *Vilnis et al.*
- Goal: representing entities as high-dimensional products-of-intervals (e.g. hyperrectangles, also called *boxes*), where the event's unary probability comes from the box volume and the joint probabilities of events come from the boxes' intersections



Probabilistic Box Embeddings

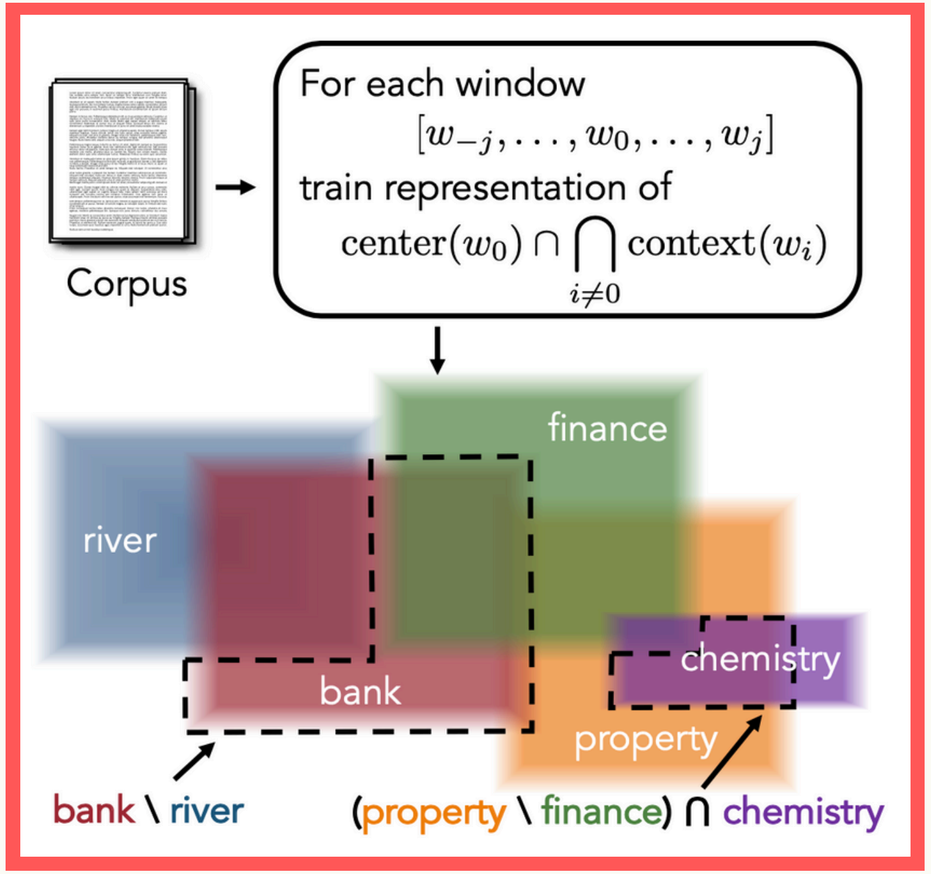


- able to capture **anticorrelation** and disjoint concepts ✓
- able to perform rich joint and conditional queries over arbitrary sets of concepts ✓
- able to learn from and predict calibrated **uncertainty** ✓
- not all concepts can be represented as boxes (for example the complementary of a box is not a box) ✗
- hard to assure that the total probability of the boxes is equal to 1 ✗
- learning is not straightforward ✗

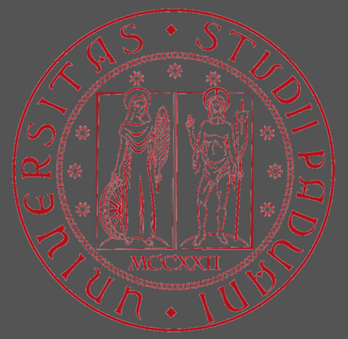


WORD2BOX

- A **geometric word embedding** algorithm proposed in 2022 by Dasgupta *et al.*
- a fuzzy set interpretation of box embeddings for words
- based on a variant of box embeddings called **GumbelBox**
- set-theoretic training objective allows to capture relationships between words that cannot be grasped with traditional methods



Box Embeddings



The box embeddings of an element *a* is a Cartesian product of intervals:

$$Box(\mathbf{x}) := \prod_{i=1}^{d} [x_i^-, x_i^+] = [x_1^-, x_1^+] \times \ldots \times [x_d^-, x_d^+] \subseteq \mathbb{R}^d$$

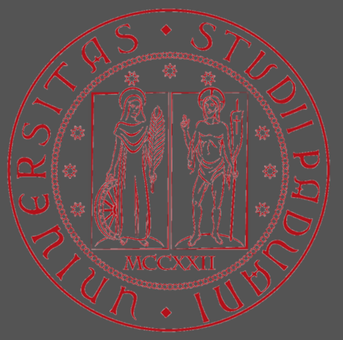
The **unary probability** of an event *a* is defined as the volume of the box *Box(x)* associated with it:

$$P(a) = |Box(\mathbf{x})| = \prod_{i=1}^{d} max(0, x_i^+ - x_i^-)$$

The **joint probability** of two events *a*, *b* is defined as the volume of the intersection of their boxes, i. e. Box(x) and Box(y) respectively:

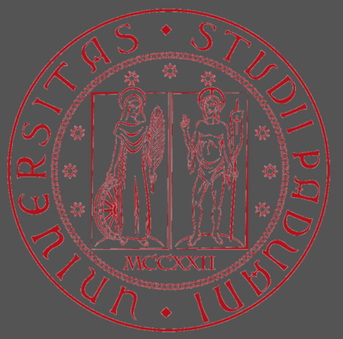
$$P(a,b) = |Box(\mathbf{x}) \cap Box(\mathbf{y})| = |\prod_{i=1}^{d} [max(x_i^-, y_i^-), min(x_i^+, y_i^+)]|$$

Gumbel Boxes



PROBLEM: hard *max* and *min* operations are not differentiable, resulting in large areas of the parameter space where the gradient is zero

Gumbel Boxes



PROBLEM: hard *max* and *min* operations are not differentiable, resulting in large areas of the parameter space where the gradient is zero

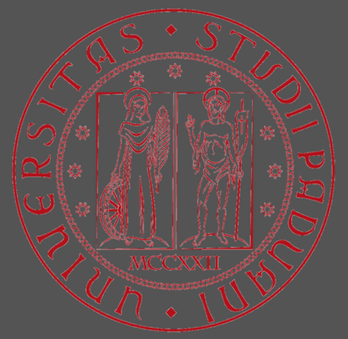
SOLUTION: model the corners of the boxes with Gumbel random variables

Gumbel Boxes

Advantages

- the Gumbel distribution is min/max stable, meaning that the minimum and maximum of two Gumbel random variables are also Gumbel distributed ➡ the intersection between two Gumbel boxes is still a Gumbel box
- the GumbelBox method outperforms other variants of box embeddings
- Gumbel boxes embedded in a space of finite measure have a **rigorous probabilistic interpretation**

Fuzzy sets



Definition

- generalization of the classical sets
- elements have degrees of membership
- **membership function** m : U → [0, 1]
- An element x is called:
  - not included in (U, m) if m(x) = 0
  - fully included in (U, m) if m(x) = 1
  - partially included in (U, m) if 0 < m(x) < 1

Fuzzy sets

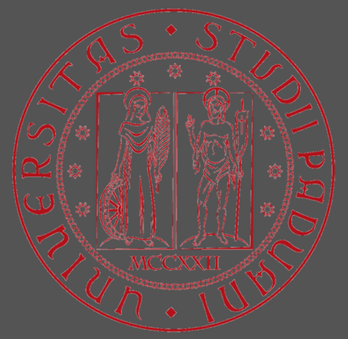
Definition

- generalization of the classical sets
- elements have degrees of membership
- **membership function** m : U → [0, 1]
- An element x is called:
  - not included in (U, m) if m(x) = 0
  - fully included in (U, m) if m(x) = 1
  - partially included in (U, m) if 0 < m(x) < 1

Reasons

- not possible to learn a set representation in a gradient-based model using a hard membership function
- able to model **graded similarity**, i.e. the capabilty of measuring the degree of similarity between entities along a continuum or a scale

Gumbel Boxes as Fuzzy Sets



A Gumbel box Box(x), x ∈ R^2d, corresponds to the fuzzy set (R^d, m) where m : Rd → [0, 1] is defined as the probability of a point z ∈ Rd being inside the Gumbel box:

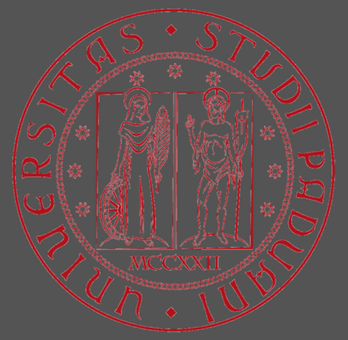
$$m(\mathbf{z}) = P(\mathbf{z} \in Box_G(\mathbf{x})) = \prod_{i=1}^{d} P(z_i > X_i^-) P(z_i < X_i^+)$$

where {X±} are Gumbel random variables that define the corners of the box.

The volume of the box is obtained by integrating its membership function:

$$|Box_F(\mathbf{x})| = \int_{\mathbb{R}^d} P(\mathbf{z} \in Box_G(\mathbf{x})) d\mathbf{z} \approx \prod_{i=1}^{d} \beta log(1 + exp(\frac{\mu_i^+ - \mu_i^-}{\beta} - 2\gamma))$$

Model Training



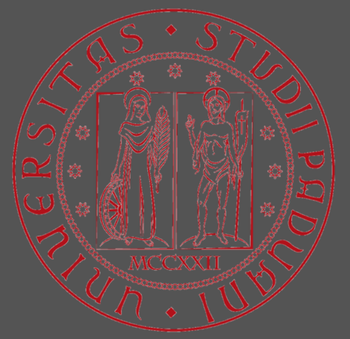
Goal: to learn the following center and context box representations, based on fuzzy sets, for each word *v* in the vocabulary *V*

$$cen_B(v) := Box_F(\widetilde{cen_W}(v))$$
$$con_B(v) := Box_F(\widetilde{con_W}(v))$$

using a **max-margin** training objective where the score for a given window **w** is defined as:

$$f(\mathbf{w}) := \left| cen_B(w_0) \cap \bigcap_{i \neq 0} con_B(w_i) \right|$$

Model Training



Goal: to learn the following center and context box representations, based on fuzzy sets, for each word *v* in the vocabulary *V*

$$cen_B(v) := Box_F(\widetilde{cen_W}(v))$$
$$con_B(v) := Box_F(\widetilde{con_W}(v))$$

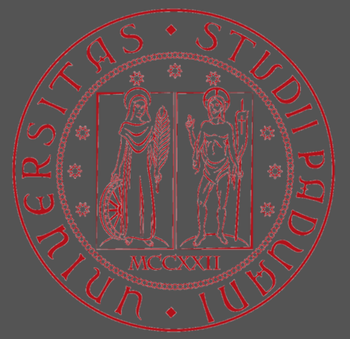
using a **max-margin** training objective where the score for a given window **w** is defined as:

$$f(\mathbf{w}) := \left| cen_B(w_0) \cap \bigcap_{i \neq 0} con_B(w_i) \right|$$

Dataset

- preprocessed **ukWaC** corpus (more than 112k unique words)
- negative examples + subsampling of context words

Model Training



Goal: to learn the following center and context box representations, based on fuzzy sets, for each word *v* in the vocabulary *V*

$$cen_B(v) := Box_F(\widetilde{cen_W}(v))$$
$$con_B(v) := Box_F(\widetilde{con_W}(v))$$

using a **max-margin** training objective where the score for a given window **w** is defined as:

$$f(\mathbf{w}) := \left| cen_B(w_0) \cap \bigcap_{i \neq 0} con_B(w_i) \right|$$

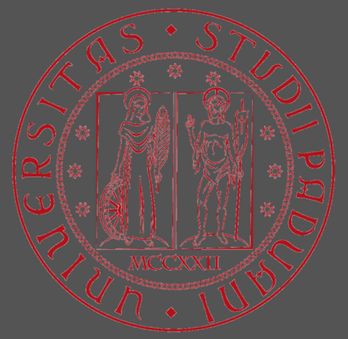
Dataset

- preprocessed **ukWaC** corpus (more than 112k unique words)
- negative examples + subsampling of context words

Baseline

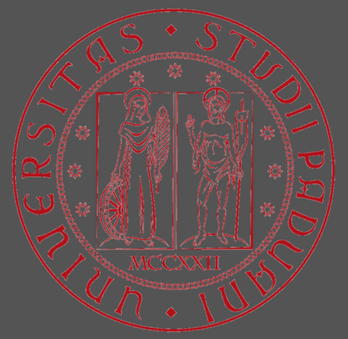
128-dimensional Word2Vec model trained using **CBOW with negative sampling**

Model Evaluation



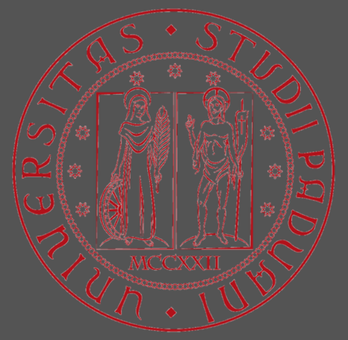
1. Word Similarity Benchmark
    - datasets made of word pairs (both nouns and verbs) annotated by humans with a similarity score
2. Set Theoretic Operations
    - dataset of homographs and polysemic words, consisting of triples of words (A, B, C) where A ◦ B should yield a set similar to C, for some set-theoretic operations ◦

Model Evaluation



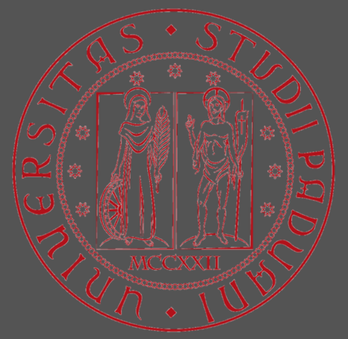
1. Word Similarity Benchmark
   - datasets made of word pairs (both nouns and verbs) annotated by humans with a similarity score
2. Set Theoretic Operations
   - dataset of homographs and polysemic words, consisting of triples of words (A, B, C) where A ◦ B should yield a set similar to C, for some set-theoretic operations ◦

     - Word2Box outperformed Word2Vec, especially on **rare words** datasets
     - more flexible representation of words
     - better consistency in the results of logical queries



Dataset

- an extract of the **enwik8** dataset created by Matt Mahoney
- data preprocessing
  - XML tags and external urls are removed
  - numbers are replaced with the <NUM> token
  - punctuation marks are replaced with the corresponding token
  - stopwords are removed
  - low-frequency words are removed
- data exploratory analysis



Dataset

- an extract of the **enwik8** dataset created by Matt Mahoney
- data preprocessing
  - XML tags and external urls are removed
  - numbers are replaced with the <NUM> token
  - punctuation marks are replaced with the corresponding token
  - stopwords are removed
  - low-frequency words are removed
- data exploratory analysis

<u>Resulting Dataset</u>

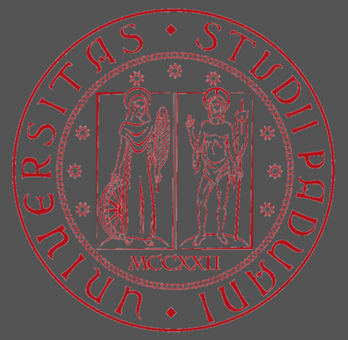
- 8 858 098 words
- 53 693 unique words

MY EXPERIMENTS

Before the preprocessing...

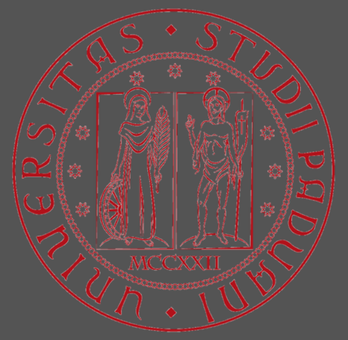
```
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.3/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" x
si:schemaLocation="http://www.mediawiki.org/xml/export-0.3/ http://www.mediawiki.org/xml/export-0.3.xsd" version="0
.3" xml:lang="en">
<siteinfo>
  <sitename>Wikipedia</sitename>
  <base>http://en.wikipedia.org/wiki/Main_Page
</base>
  <generator>MediaWiki 1.6alpha</generator>
  <case>first-letter</case>
    <namespaces>
    <names
pace key="-2">Media</namespace>
    <namespace key="-1">Special</namespace>
    <namespace key="0" />
    <na
mespace key="1">Talk</namespace>
    <namespace key="2">User</namespace>
    <namespace key="3">User talk</name
space>
    <namespace key="4">Wikipedia</namespace>
    <namespace key="5">Wikipedia talk</namespace>
    <na
mespace key="6">Image</namespace>
    <namespace key="7">Image talk</namespace>
    <namespace key="8">MediaWik
i</namespace>
    <namespace key="9">MediaWiki talk</namespace>
    <namespace key="10">Template</namespace>

     <namespace key="11">Template talk</namespace>
    <namespace key="12">Help</namespace>
    <namespace key="
```

MY EXPERIMENTS



...after the preprocessing

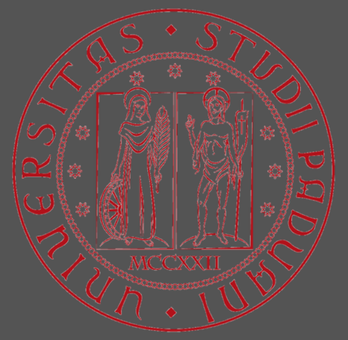
```
<QUOTATION_MARK> anarchism <QUOTATION_MARK> originated term abuse first used early working class radicals including
 diggers english revolution <QUOTATION_MARK> sans <QUOTATION_MARK> french revolution <PERIOD> whilst term still use
d pejorative way describe <QUOTATION_MARK> act used violent means destroy organization society <QUOTATION_MARK> <CO
MMA> taken positive label self defined anarchists <PERIOD> word <QUOTATION_MARK> anarchism <QUOTATION_MARK> derived
 greek <QUOTATION_MARK> <QUOTATION_MARK> <LEFT_PAREN> without archons <LEFT_PAREN> ruler <COMMA> chief <COMMA> king
 <RIGHT_PAREN> <RIGHT_PAREN> <PERIOD> anarchism political philosophy <COMMA> belief <QUOTATION_MARK> rulers <QUOTAT
ION_MARK> unnecessary abolished <COMMA> although differing interpretations means <PERIOD> anarchism refers related
social movements <RIGHT_PAREN> advocate elimination authoritarian institutions <COMMA> particularly state <PERIOD>
word anarchy <COMMA> anarchists use <COMMA> imply chaos <COMMA> nihilism <COMMA> <COMMA> rather harmonious anti aut
horitarian society <PERIOD> place regarded authoritarian political structures coercive economic institutions <COMMA
```



Word2Box code analysis

Original codebase

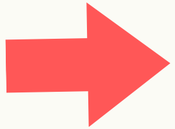
- several bugs
- not well documented



Word2Box code analysis

Original codebase

- several bugs
- not well documented



Colab Notebook

- bugfixing
- detailed documentation
- ready-to-run training and model's evaluation

FUTURE WORK

- train the Word2Box on the full dataset
- better tuning of the hyperparameters
- evaluate the model's performances in downstream NLP tasks
- comparison with the static word embedding model proposed by Ahmet Onur Akman in his Master's Thesis *"Design of a Third-Order Word Embedding Model Using Vector Projections" (2023)*