KATHOLIEKE UNIVERSITEIT
LEUVEN

FACULTEIT
INGENIEURSWETENSCHAPPEN

Master
Computer Science

Milan van der Meer

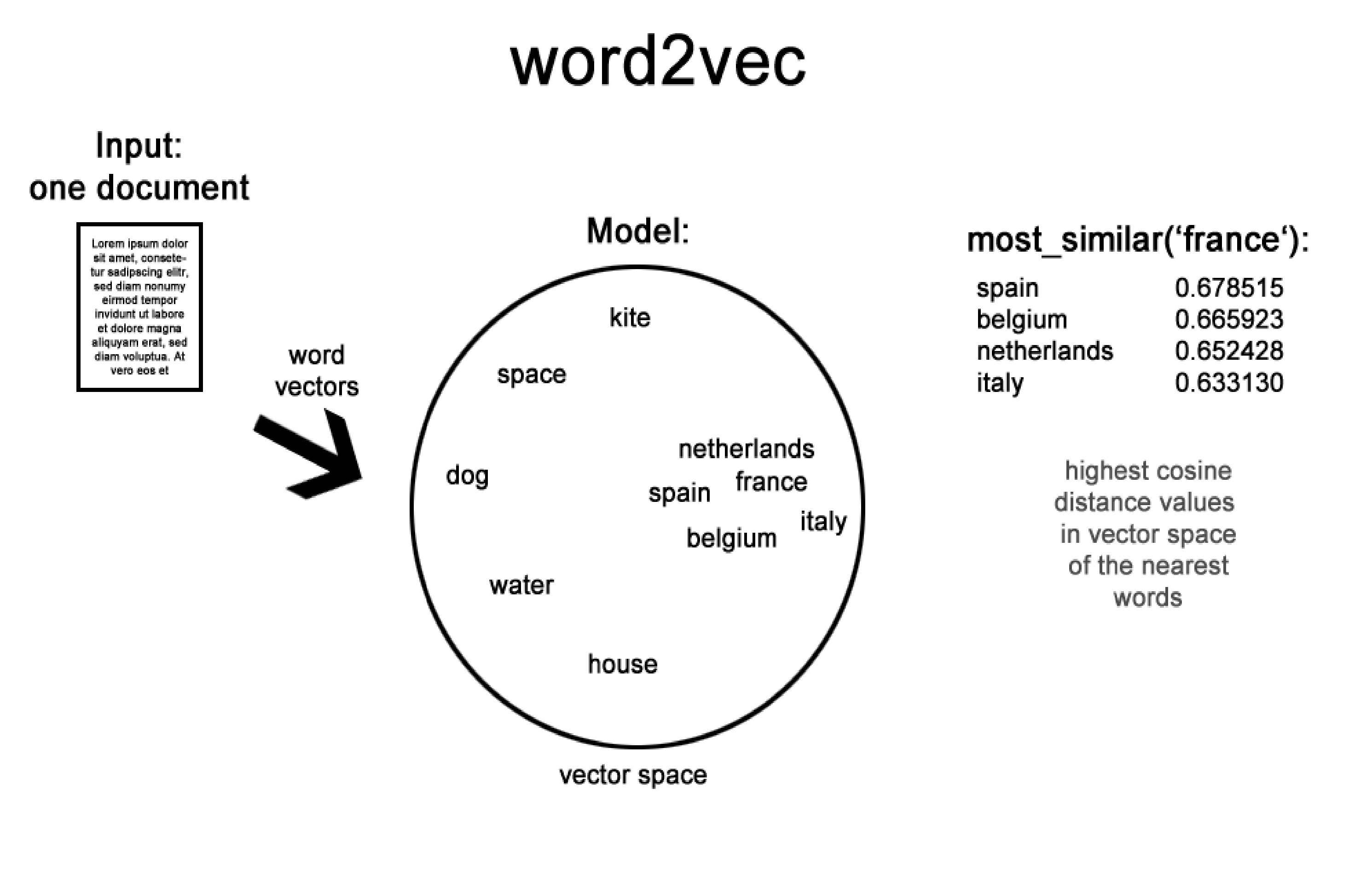
Prof. dr. R. Wuyts and A. Vapirev

Academic year 2015-2016

Learning a Disease Embedding using Generalized Word2Vec Approaches

Electronic Health Records (EHR)

- Personal medical data
  - Doctor visits
  - Lab results
  - Demographics
- Increased usage of EHRs
- Lots of potential


word2vec
Input:
one document
Lorem ipsum dolor sit amet, consete-tur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et
word vectors
Model:
kite
space
dog
netherlands
spain
france
italy
belgium
water
house
vector space
most_similar('france'):
spain 0.678515
belgium 0.665923
netherlands 0.652428
italy 0.633130
highest cosine distance values in vector space of the nearest words

Generalized Word2Vec

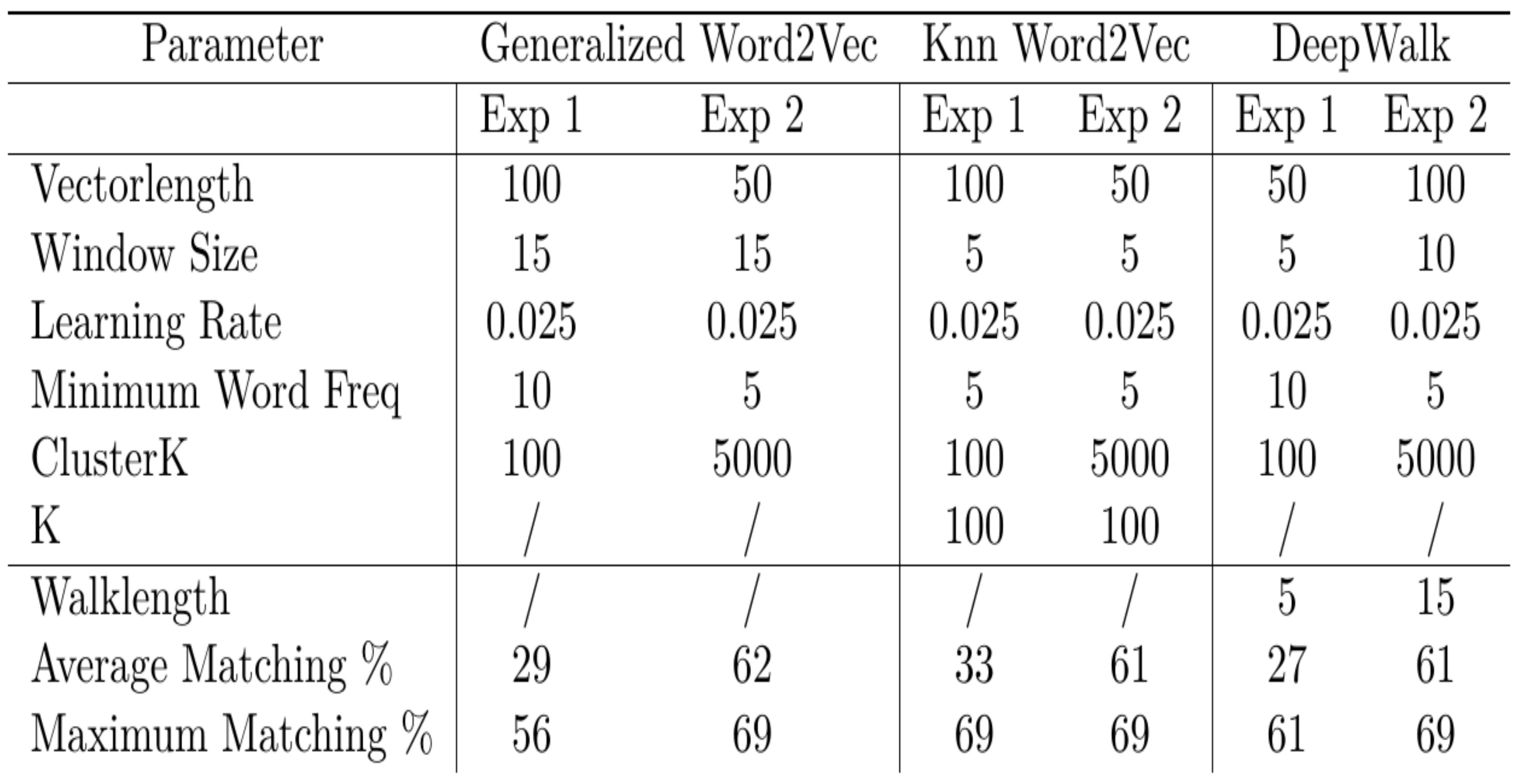
- Analogy between sentences of words and sequences of EHR events
- 3 proposed methods
  - Generalized Word2Vec
  - Knn Word2Vec
  - Generalized DeepWalk
- Find relations between diseases using Generalized Word2Vec
- Handle new EHR events with k-nearest neihbor methods
- Make performant with DeepWalk

EHR Analytics

- New research area
- Problems
  - Privacy
  - Different codings
- Goals
  - Find disease trajectories
  - Test drug treatments
- Methods
  - Querying
  - Statistics
  - Out-of-the box machine learning
- Generalized Word2Vec

Results

- Validate using Danish paper
- Compare generated Word2Vec clusters with Danish clusters
- Basic parameter tuning
- Conclusion
  - Clusters match well enough
  - Especially with estimations taken into account



Parameter	Generalized Word2Vec		Knn Word2Vec		DeepWalk	
	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2
Vectorlength	100	50	100	50	50	100
Window Size	15	15	5	5	5	10
Learning Rate	0.025	0.025	0.025	0.025	0.025	0.025
Minimum Word Freq	10	5	5	5	10	5
ClusterK	100	5000	100	5000	100	5000
K	/	/	100	100	/	/
Walklength	/	/	/	/	5	15
Average Matching %	29	62	33	61	27	61
Maximum Matching %	56	69	69	69	61	69