Semantic Embedding for Information Retrieval

Shenghui Wang, Rob Koopman
OCLC Research, Leiden, The Netherlands
{shenghui.wang, rob.koopman}@oclc.org

A word is characterised by the company it keeps

Words that occur in similar contexts tend to have similar meanings

Word embedding approaches
- Count-based methods (e.g. LSA, Random Projection)
- Predictive methods (e.g. neural probabilistic language models)

From word to document distance
- Bag of Words or TF-IDF
- LSI, LDA
- Weighted average of word embeddings
- Neural network based methods, e.g. Doc2Vec
- Word mover's distance

Dataset
- Metadata of 27 million Medline articles
- Fields: title, abstract, subject, author, affiliation, journal, citation, etc.
- Each article is a sequence of title words and entity tokens (e.g. subject:eczema physiology and author:diefenbach wc)

Embedding methods to compare
- Word2Vec / Doc2Vec
- GloVe
- Ariadne

Evaluation I: Word analogy test

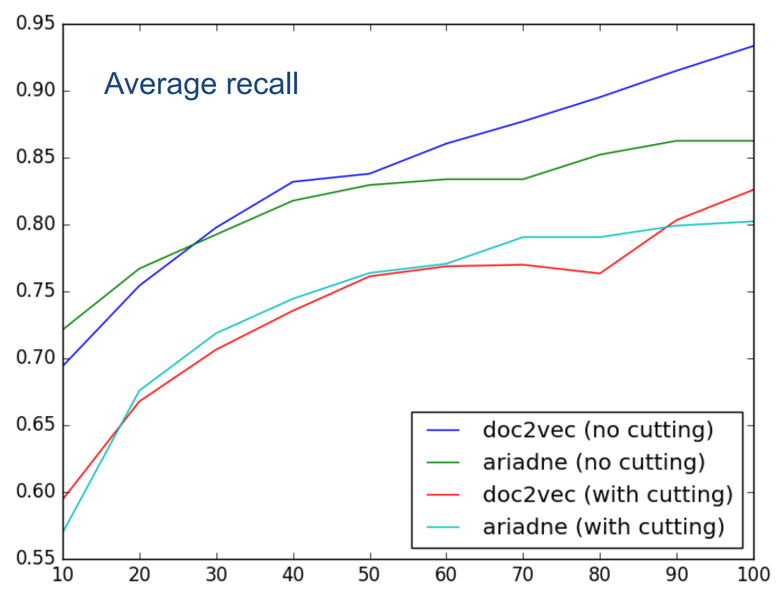
vector("Paris")-vector("France") + vector("Italy") ≈ vector("Rome")?

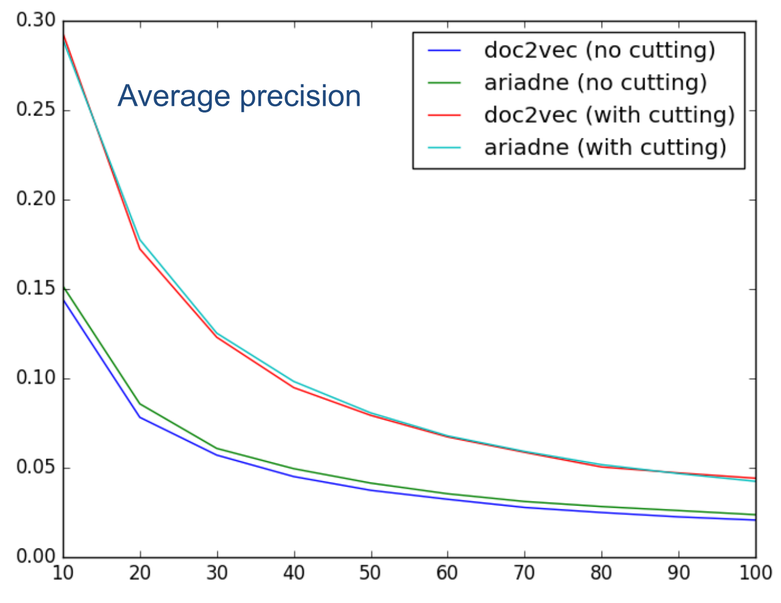
Method	Accuracy(%)	Training time (seconds)	# Thread
Ariadne	1.6	15,020	1
Word2Vec	62.7	38,364	16
GloVe	53.6	22,680	16

Different embedding methods lead to different word embeddings

Word	Method	Top 10 most similar words
frog	Ariadne	sartorius, frogs, rana, liagushki, temporaria, liagushek, catesbiana, sartorii, amphibian, caudiverbera
	Word2Vec	toad, bullfrog, amphibian, rana, frogs, turtle, bufo, salamander, caudiverbera, newt
	GloVe	rana, frogs, amphibian, toad, temporaria, bullfrog, laevis, xenopus, ridibunda tadpoles
brain	Ariadne	brains, cortical, cortex, forebrain, cerebellum, neocortex, neuronal, neuroanatomical, neural, limbic
	Word2Vec	cerebral, cerebellum, cns, brains, brainstem, hippocampus, forebrain, cerebrum, cortical, neocortex
	GloVe	cerebral, brains, cns, nervous, neuronal, cerebellum, hippocampus, neurological, cortex, cerebrum
knee	Ariadne	knees, tibiofemoral, femorotibial, tibial, kneeling, joint, malalignment, flexion, unicompartmental, tka
	Word2Vec	hip, ankle, elbow, knees, shoulder, joint, patellofemoral, wrist, patellar, acl
	GloVe	knees, joint, hip, ankle, osteoarthritis, arthroplasty, joints, cruciate, elbow, flexion
depression	Ariadne	depressive, mood, nondepressed, subsyndromal, depressed, anxiety, dysthymia, phq, hamilton, anxious
	Word2Vec	depressive, anxiety, insomnia, mdd, psychopathology, psychosis, ptsd, mood, suicidality, mania
	GloVe	depressive, anxiety, depressed, mood, psychiatric, symptomatology, psychological, affective, psychopathology, emotional
insulin	Ariadne	hyperinsulinemia, glucose, hyperglycemia, insulinopenia, eug-lycemia, normoglycemic, hypoinsulinemia, insulinemia, nondiabetic, glycemia
	Word2Vec	glucagon, gh, leptin, glucose, gip, hyperinsulinemia, adiponectin, niddm, glp, hyperinsulinemic
	GloVe	glucose, diabetes, glucagon, mellitus, fasting, hyperglycemia, leptin, igf, diabetic, hyperinsulinemia
treatment	Ariadne	treated, treat, therapy, treating, efficacy, discontinued, received, discontinuation, clinical, option
	Word2Vec	therapy, treatments, treating, monotherapy, pharmacotherapy, management, chemotherapy, prophylaxis, intervention, therapeutic
	GloVe	treated, treatments, therapy, treating, therapeutic, effective, further, treat, with, results
vitamin	Ariadne	vitamins, vit, hydroxyvitamin, hypovitaminosis, vitd, cholecalciferol, calcidiol, supplements, supplementation, ergocalciferol
	Word2Vec	vitamins, vit, vitamine, hypovitaminosis, hydroxyvitamin, avitaminosis, cholecalciferol, vitamina, folate, selenium
	GloVe	vitamins, supplementation, dietary, folic, tocopherol, supplements, ascorbic, deficiency, hydroxyvitamin, d3

Evaluation II: Updating medical guidelines
- 29 statements (16 breast caner, 4 hepatitis C, 4 lung cancer, 5 ovarian cancer)
- 103 source articles, 156 target articles, in total 180 unique articles
- 1 million English articles, with abstract, published between 1984 and 2012





Predict or count smartly

OCLC