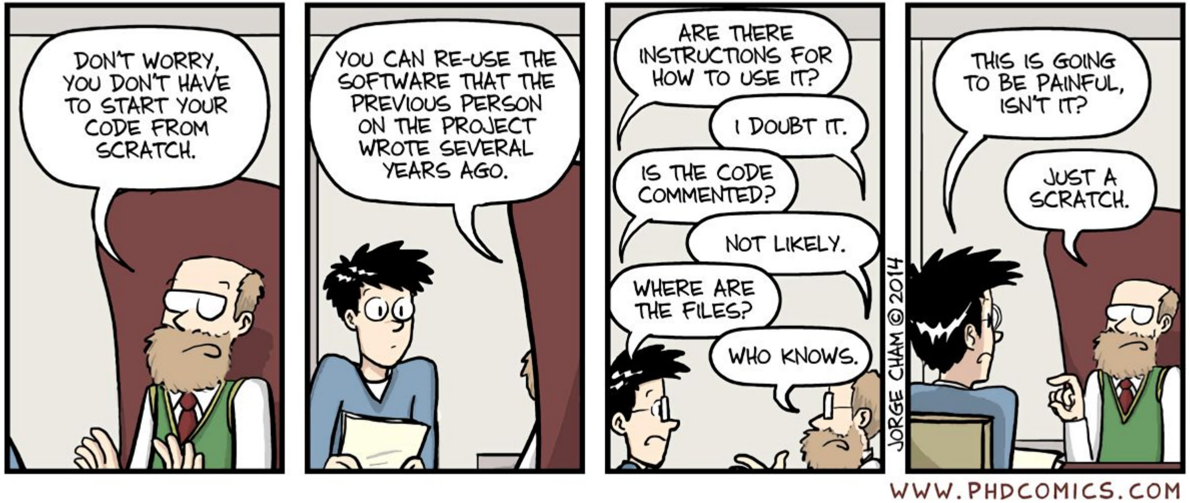
The Art of Reproducible Machine Learning

A Survey of Methodology in Word Vector Experiments


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM

Vítek Novotný, witiko@mail.muni.cz

https://mir.fi.muni.cz/

RASLAN 2020

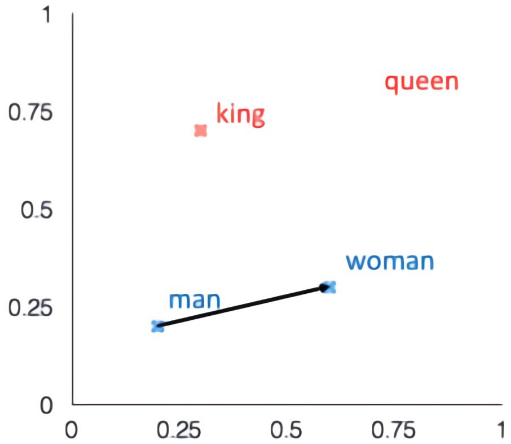
Word Analogy

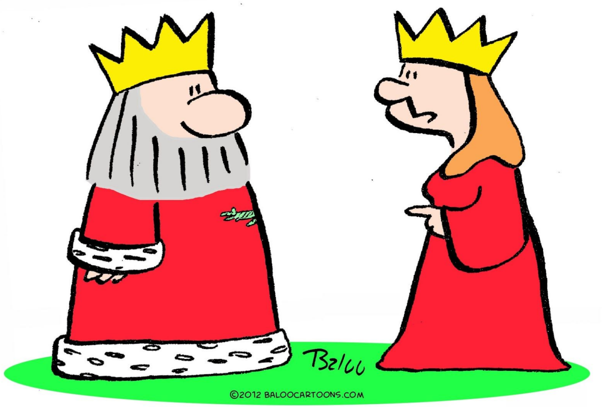
- **Word analogy** [5] measures how well word vectors can answer the question

"Which word b' is to a' as a is to b?"

man:woman :: king:?

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





"Your new robe is nice, but I don't like the little alligator."

Solution b' = queen for a = man, b = woman, a' = king

[5]: arxiv.org/pdf/1301.3781.pdf (Efficient Estimation of Word Representations in Vector Space)

Word Analogy Limiting and Caseless Matching

- In word analogy, we only use the *N* most frequent words as candidates for *b’*.
- *N* is either undisclosed [1–3], or it ranges from $2 \cdot 10^5$ [4] to $1 \cdot 10^6$ [5].
- Reproduce Grave [4] with different *N*’s, get *16% difference in accuracy*.

- In word analogy, we must find the words *a*, *b*, *a'*, *b'* in the vector vocabulary.
- Some implementations use upper-casing, some lower-casing, some neither.
- In Unicode, case is neither bijective nor transitive, and is locale-dependent:
  - Upper-casing maps ß to SS, and lower-casing maps SS to ss (not ß).
  - Lower-casing maps I to ı in Turkish and Azari, and to i in other locales.
- Reproduce Grave with different locales and cases, get *18% diff. in accuracy*.

[1]: arxiv.org/pdf/1310.4546.pdf (Distributed Representations of Words and Phrases and their Compositionality)
[2]: www.aclweb.org/anthology/Q17-1010.pdf (Enriching Word Vectors with Subword Information)
[3]: www.lrec-conf.org/proceedings/lrec2018/pdf/721.pdf (Advances in Pre-Training Distributed Word Representations)
[4]: arxiv.org/pdf/1802.06893.pdf (Learning Word Vectors for 157 Languages)
[5]: arxiv.org/pdf/1301.3781.pdf (Efficient Estimation of Word Representations in Vector Space)

Multi-Word Expressions

- **Phrasing algorithm** [1] merges common word bigrams w_i, w_j into phrases:

$$\mathrm{score}(w_i, w_j) = \frac{\mathrm{count}(w_i w_j)}{\mathrm{count}(w_i) \cdot \mathrm{count}(w_j)}$$

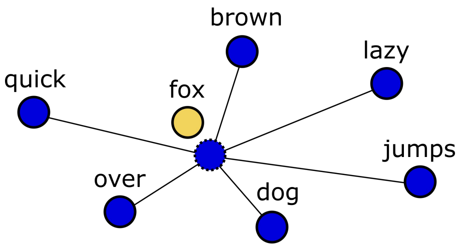
- Mikolov [1] merge bigrams w_i, w_j when **score**(w_i, w_j) > δ, but don't disclose δ.
- Mikolov [1] repeat merging to form longer phrases with undisclosed decay of δ.
- Reference implementation and Gensim implementation both differ from **score**.
- Reference implementation and Gensim implementation both use different δ.
- Reference implementation only uses $N = 5 \cdot 10^8$ most frequent words for w_i, w_j.
- We failed to reproduce [6] *any* increase in English word analogy accuracy.

[1]: arxiv.org/pdf/1310.4546.pdf (Distributed Representations of Words and Phrases and Their Compositionality)

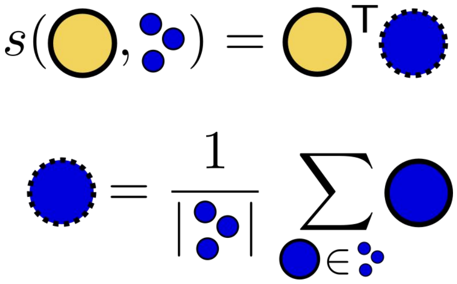
[6]: arxiv.org/pdf/1712.09405.pdf (Advances in Pre-Training Distributed Word Representations)

Positional Weighting

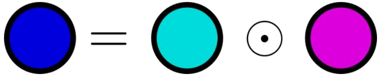
- **Baseline model** predicts a masked word from the mean context word vector:



“The quick brown **???** jumps over the lazy dog”


$s($, $) =$ $^{\mathsf{T}}$
$=$ $\frac{1}{|\;|}$ $\sum$ $\in$

- **Positional model** [6, 2.2] makes context word vectors depend on position:
  - Context “Unlike dogs, cats are **???**” has a different vector than “Unlike cats, dogs are **???**”.
  - Mikolov et al. [6] do not disclose the initialization of context and position vectors.
  - Try different init.’s with 2017 English Wikipedia [8], get *24% difference in word analogy accuracy*.

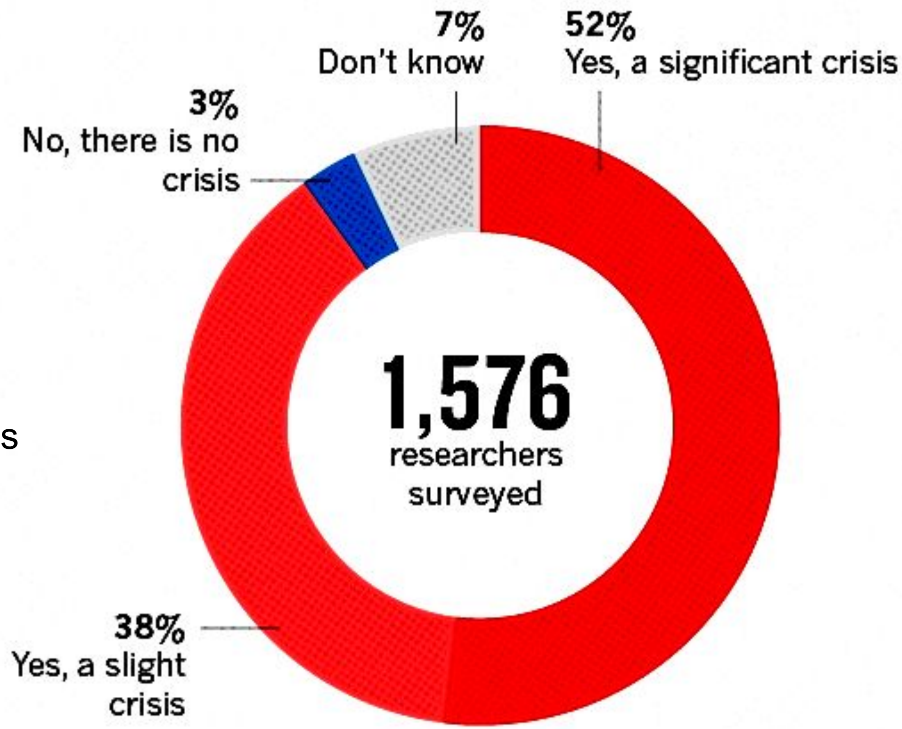


[6]: arxiv.org/pdf/1712.09405.pdf (Advances in Pre-Training Distributed Word Representations)

[8]: github.com/RaRe-Technologies/gensim-data/releases/tag/wiki-english-20171001

Conclusion Is There a Reproducibility Crisis? [7]

- Many factors contribute to the crisis:
  1. *Rapid research* in machine learning
  2. *Publish-or-perish* in academia
  3. Ever-increasing *model complexity*
- *Reproducibility* and *comparability* depend on controlling *all* variables.
- We hope that our study will:
  1. Make it *easier to reproduce* both previous and future word vector experiments
  2. Serve as an *inspiration* for upholding the principles of reproducibility in future machine learning research
- Thank you for your attention!



[7]: nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970 (1,500 scientists lift the lid on reproducibility)