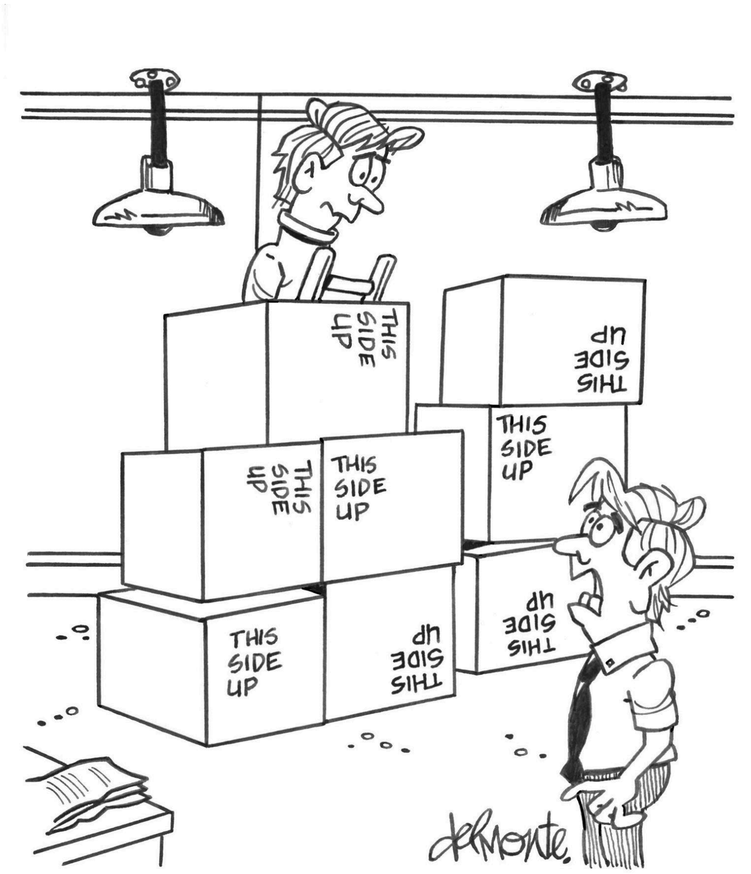
Quiz 9: Topic Modeling for Fun and Profit



"Come down here, Wilson, we need to talk about your reading comprehension."

CartoonStock.com

Simple English Wikipedia is a subset of Wikipedia for

- Students
- Children
- Teachers
- Adults with learning disabilities
- Non-native English speakers

and is written in a language that is easy to understand but is still natural and grammatical. The articles in the Simple English Wikipedia use shorter sentences and easier words and grammar than the regular English Wikipedia. This makes articles easier to understand and change. It has (only!) 250,222 articles, compared with 6,804,482 for its big brother.

Radim Řehůřek is the author of a popular LDA implementation (Gensim). His post on Gensim is available online, targeted at SimpleWiki. The code presented therein doesn't work but this Jupyter notebook, a close cousin, does! That will be our starting point! The original and the fixed-up versions of the code are not identical, however.

The example notebook runs under Google Colab, not under Spark. **Runtime | Run all** in the notebook takes about 35 minutes – you don't want to do it too often!

The notebook has 2 questions embedded in it. Please answer them.

1. [6 points] Print all words and their ids from `id2word_wiki` where the word starts with "human".
2. [7 points] Print `text` transformed into TFIDF space.

You will also notice that cell #37 of the fixed-up notebook `# select top 50 words for each of the 20 LDA topics` produces answers that are very different from what cell #30 of Radim's notebook *# select top 50 words for each of the 20 LDA topics* produces.

3. [14 points] Identify the source of difference and change it so they are equivalent[1].

Topic Tagging

One challenge of LDA is naming the topics. To ease this work, the techniques mentioned in the notebook can help. Use each to identify the topic.

4. [8 points] Misplaced word technique
5. [8 points] Half & half technique: split each document into two parts, and check that 1) topics of the first half are similar to topics of the second 2) halves of different documents are mostly dissimilar.
6. [7 points] Which algorithm, LSI or LDA, performs better for this dataset. Please justify your answer.

[1] They still won't be identical because Radim's post was written some years ago, when Simplewiki was smaller.