Large-Scale Data Processing and Data Exploration with R: Project Task Description

Apl. Prof. Dr. Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

April 30, 2020

1 Idea

You are provided a selection of norms for English and German across a range of variables. The norms rely on human judgements and/or semi-automatic extensions regarding degrees of concreteness, valence, arousal, imageability and further perception modalities. In addition, you are provided corpus-based frequency lists as well as distributional co-occurrence scores.

The goal of your project is to first analyse a subset of the norm data and then to explore whether judgements are related across modalities and to corpus-based frequency and semantic diversity.

2 Data

You are provided the following datasets.

1. **English Norms**
   - ***extension of PYM concreteness, imagery and meaningfulness norms*** (Paivio et al., 1968; Clark and Paivio, 2004)
   - ***MRC database*** (Coltheart, 1981)
   - ***modality exclusivity norms*** (Lynott and Connell, 2009, 2013)
   - ***concreteness norms*** (Turney et al., 2011)
   - ***ANEW valence, arousal and dominance norms*** (Warriner et al., 2013)
   - ***concreteness norms*** (Brysbaert et al., 2014)
   - ***NRC-VAD valence, arousal and dominance norms*** with translations into 100 languages (Mohammad, 2018)

2. **German Norms**

   - ***concreteness, valence and arousal norms*** (Lahl et al., 2009)
   - ***Leipzig affective norms*** (Kanske and Kotz, 2010)
   - ***ANGST affective norms for sentiment terms*** (Schmidtke et al., 2014)
   - ***concreteness, imageability, valency and arousal norms***
     (Köper and Schulte im Walde, 2016)

3. **Corpus Frequencies**

   - `encow16-freqs.txt.gz`: a gzipped corpus frequency text file for English with three tab-separated columns (lemma, part-of-speech, frequency),
     see `http://corporafromtheweb.org/encow16/` for details
   - `decow16-freqs.txt.gz`: a gzipped corpus frequency text file for German with three tab-separated columns (lemma, part-of-speech, frequency),
     see `http://corporafromtheweb.org/decow16/` for details

   The COW corpora are described in (Schäfer and Bildhauer, 2012; Schäfer, 2015).

4. **Distributional Information**

   - `encow16-window-2_freqs_lmi.txt.gz`:
     a gzipped file for co-occurrence of lemma/part-of-speech target–context pairs within a window of 2 words (left + right);
     four tab-separated columns (target:::pos, context::pos, frequency, lmi score)
   - `decow16-window-2_freqs_lmi.txt.gz`:
     a gzipped file for co-occurrence of lemma/part-of-speech target–context pairs within a window of 2 words to the (left + right);
     four tab-separated columns (target:::pos, context::pos, frequency, lmi score)

   As a general reminder to distributional information and corpus co-occurrence you can find a video in ILIAS.

   Local mutual information (lmi) scores are described on `www.collocations.de/AM/` in "Measures from Information Theory"; also see Evert (2009).

The norm data including references to articles are available from ILIAS. The frequency and window data are available in `/mount/studenten/R-project/2020/`. In the same directory you also find a link to a work space where you will later be allowed to create your own directory to work on the project.

3 Tasks

1. Choose at least three variables either from the same norm dataset or from different norm datasets. You can focus on English or on German or perform a cross-lingual exploration taking translations into account (such as those in the NRC-VAD dataset).

2. Preprocess the norm dataset as well as the frequency and window files, in order to down-scale large-scale information in a both linguistically and statistically meaningful way. For example, focus on a specific target part-of-speech and/or focus on specific co-occurrence dimensions and/or get rid of low-frequency counts, etc.

3. Provide detailed descriptive statistics and plots for the norms of your choice.

4. Use your dataset to explore the following hypotheses:

   (a) The ratings of your chosen variables are related to each other across your target words.

   (b) The target ratings in the norms are related to target corpus frequency. For example, the more frequent a word is the more imaginable it is.

   (c) The target ratings in the norms are related to the semantic diversity of their distributional nearest neighbours.
   In this context, we define the degree of semantic diversity of a target word as the average *cosine* score of the target word and its k nearest neighbours. I.e., you first identify the k nearest neighbours of a target word, based on the window co-occurrences, and then you determine the average cosine score of the target word and its k nearest neighbours. You can choose your own k as long as $k \geq 10$.

4 Submission

Your report should be 5–8 pages long (excluding the bibliography).

Use the template `report.tex` (or an adequate .doc file).

The report should provide

- a description of the norm datasets and variables you chose,
- a summary of your preprocessing steps (without code),
- a summary of your descriptive statistics and the most meaningful plots,
- a summary of the methods you applied to explore the hypotheses, and
- a summary of your insights.

Upload your report named ⟨`YourLastName`⟩`_report.pdf` to ILIAS by June 26, 2020.

References

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods*, 64:904–911, 2014.

James M. Clark and Allan Paivio. Extensions of the Paivio, Yuille, and Madigan (1968) Norms. *Behavior Research Methods, Instruments, and Computers*, 36(3):371–383, 2004.

Max Coltheart. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A:497–505, 1981.

Stefan Evert. Corpora and Collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2 of *Handbooks of Linguistics and Communication Science*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, 2009.

Philipp Kanske and Sonja A. Kotz. Leipzig Affective Norms for German: A Reliability Study. *Behavior Research Methods*, 42(4):987–991, 2010.

Maximilian Köper and Sabine Schulte im Walde. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portoroz, Slovenia, 2016.

Olaf Lahl, Anja S. Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. Using the World-Wide Web to obtain Large-Scale Word Norms: 190,212 Ratings on a Set of 2,654 German Nouns. *Behavior Research Methods*, 41(1):13–19, 2009.

Dermot Lynott and Louise Connell. Modality Exclusivity Norms for 423 Object Properties. *Behavior Research Methods*, 41(2):558–564, 2009.

Dermot Lynott and Louise Connell. Modality Exclusivity Norms for 400 Nouns: The Relationship between Perceptual Experience and Surface Word Form. *Behavior Research Methods*, 45:516–526, 2013.

Saif M. Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology (Monograph Supplement)*, 76(1/2): 1–25, 1968.

Roland Schäfer. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany, 2015.

Roland Schäfer and Felix Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey, 2012.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. ANGST: Affective Norms for German Sentiment Terms, derived from the Affective Norms for English Words. *Behavior Research Methods*, 46:1108–1118, 2014.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK, 2011.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.