Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics

Tomasz Kuśmierczyk, Kjetil Nørvåg

{tomaszku, noervaag} @idi.ntnu.no

NTNU – Trondheim
Norwegian University of Science and Technology

Goal

Understanding associations between words and nutritional values to come up with new models and to improve practical applications effectiveness.

Contributions

(i) a study of a large-scale online food community in terms of relations between nutritional values and textual descriptions

(ii) the introduction of a new topic model combining text with several outputs (nutrient facts)

(iii) an evaluation of efficacy in discovering recipe topics and predicting nutritional values

Model Incentives

- nutritional values strongly correlate
- similar words are associated with all nutrient facts

Evaluation

Two practical applications:

1. Recipe topics identification (clustering):
   - our model vs. LDA+LM, 5 topics
   - more consistent weights (e.g., all positive/negative) of nutritional values
   - more focused (e.g., less mixing) topics
2. Prediction of nutritional values from text
   - our model vs. LDA+LM and LDA+GBT
   - SHARED - representation shared between all outputs
   - SEP - separate model for each of outputs
   - O1 - one of the nutrients is known
   - X1 - all but one nutrient are known
   - i-th output error [%]:

$$sMAPE_i = \frac{2}{|test|} \sum_{d \in test} \frac{|f_{id} - \hat{f_{id}}|}{|f_{id}| + |\hat{f_{id}}|}$$

References

[1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through Twitter. In Proc. of CHI, 2015.

[2] M. De Choudhury and S. S. Sharma. Characterizing dietary choices, nutrition, and language in food deserts via social media. In Proc. of CSCW, 2016.

[3] T. Kuśmierczyk, C. Trattner, and K. Nørvåg. Temporality in online food recipe consumption and production. In Proc. of WWW, 2015.

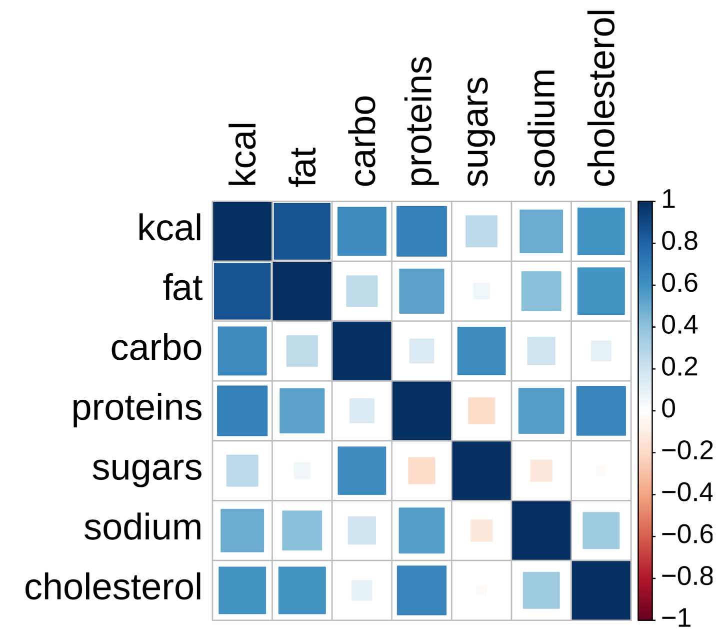
[4] D. M. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In Proc. of UAI, 2008.

[5] M. J. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In Proc. of ICWSM, 2011.

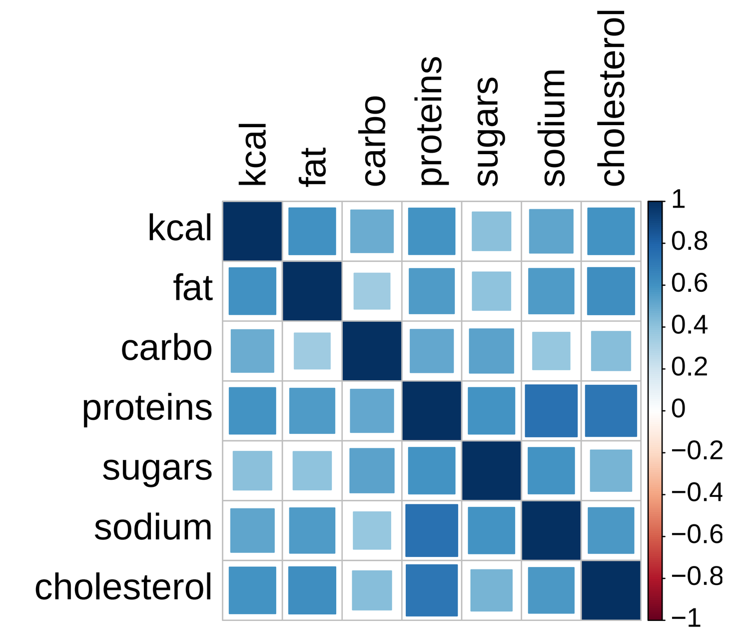
[6] R. West et al. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In Proc. of WWW, 2013.

Data set: allrecipes.com

- the largest English-language on-line cooking platform
- 58 thousand recipes
- 4,679 unique words occurring in titles at least 2 times
- 7 nutrient facts: kilocalories, fat, carbohydrates, proteins, sugars, sodium, cholesterol

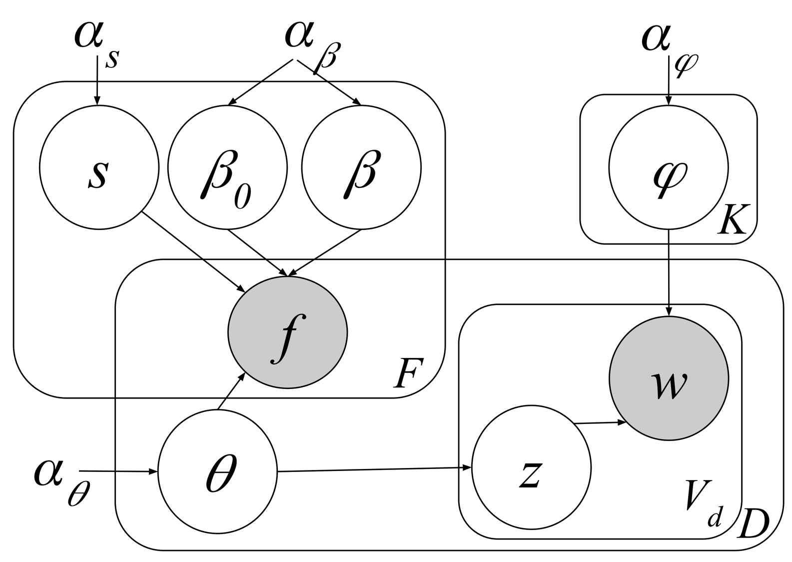


Correlations between recipe nutrients.



Correlations over words between info-gains of nutritional values.

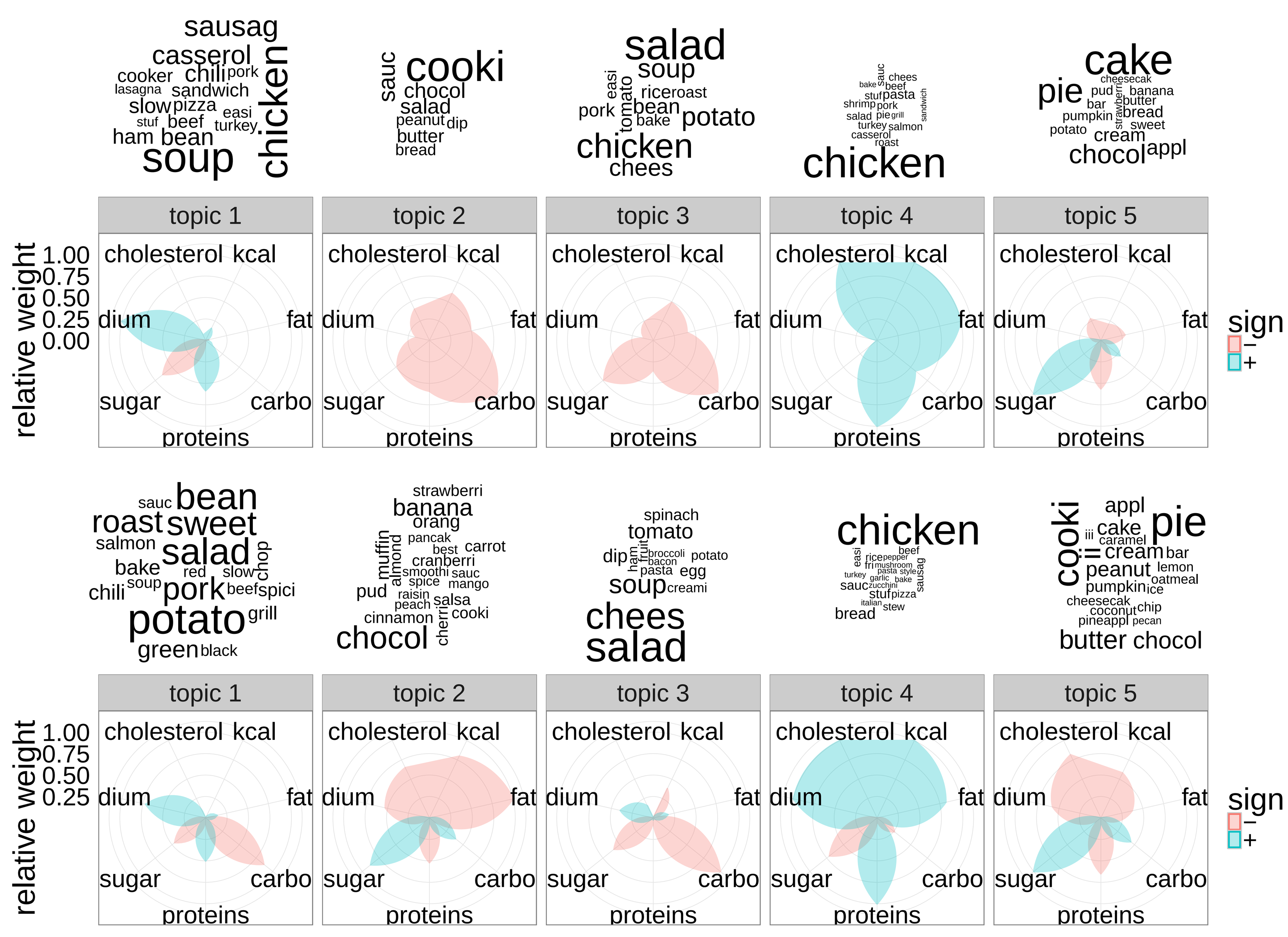
Model



K	number of topics
D	number of documents (recipes)
F	number of outputs (nutrients)
V	number of unique words
V_d	number of words in document (recipe) d
θ	the multinomial distribution of topics for recipe d
φ	word distributions for topics
z	topic assignment for word w in document d
w	observed word from document d
f_{id}	observed i-th nutrient for document (recipe) d
β_i	i-th nutrient vector of topic weights
β_0^i	i-th nutrient bias
s_i	i-th nutrient standard deviation

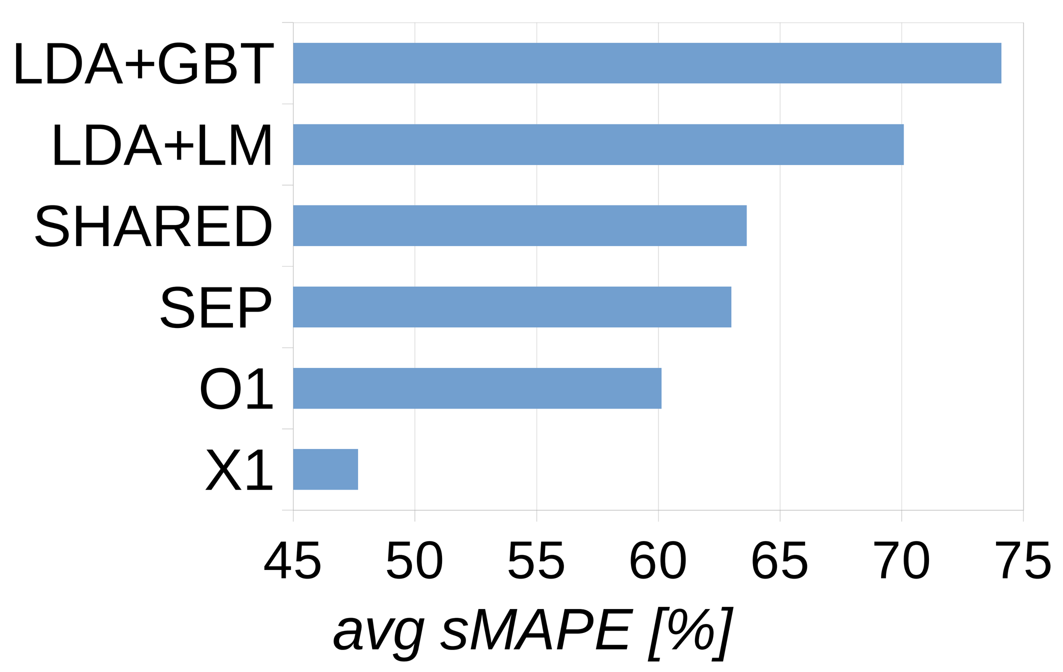
LDA with built-in multi-output linear regression extension.

Application 1: Recipe Topics Identification

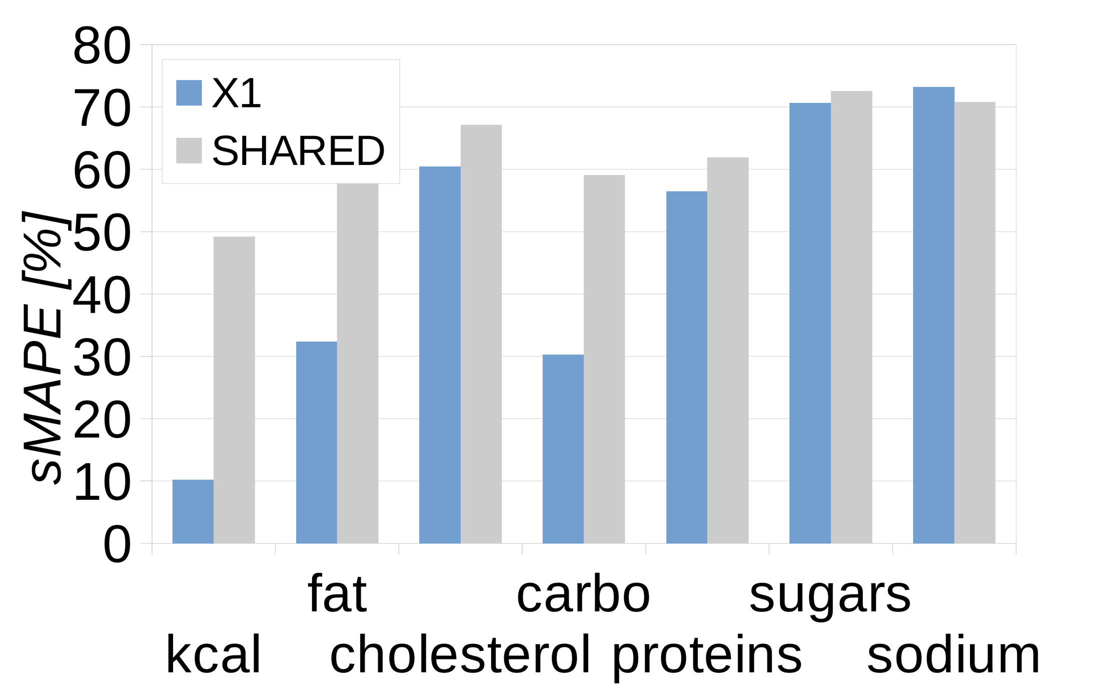


Comparison of topics found by standard LDA (bottom) and our model (top).

Application 2: Nutritional Values Prediction



Nutrients prediction avg performance.



Prediction improvement when additional outputs values are included.