Mining Dietary Supplement Product Labels

By Sakina Zabuawala

Problem Statement

The United States dietary supplement market was $41.4B in 2016. The number of supplement products in the United States increased from 4,000 in 1994 to more than 55,000 in 2012. 71% of U.S. adults—more than 170 million—take dietary supplements, according to the most recent annual survey conducted by Ipsos Public Affairs on behalf of the Council for Responsible Nutrition. 85% of U.S. adults have confidence in the safety, quality and effectiveness of dietary supplements. The top supplement categories are vitamins, minerals, specialty supplements like omega-3/fatty acids, herbals and botanicals and sports nutrition and weight management.

The US-FDA (Food and Drug Administration) regulations for dietary supplements is less restrictive compared to prescription or even over-the-counter drugs. The FDA doesn't evaluate the quality of supplements or assess their effects on the body. If a product is found to be unsafe after it reaches the market, the FDA can restrict or ban its use.

With the increasing number of supplement products in the market with little regulation and the high level of confidence in them by the consumers, there is a need to scrutinize the safety of these products.

The aim of this project is to

- Discover “topics” that group supplements with similar safety concerns
- Identify potential side-effects that might arise from specific supplements via topics
- Provide tools for pharmacovigilance and safety solutions for dietary supplements

Data

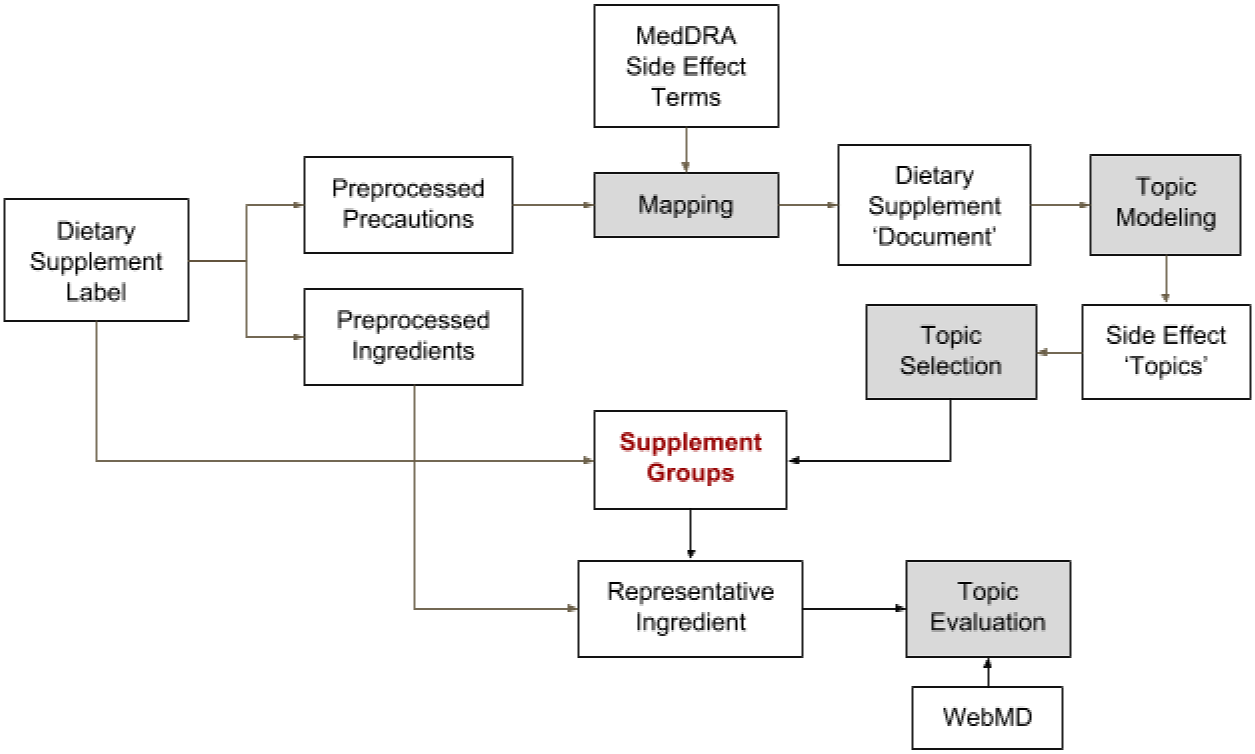
The Dietary Supplement Label Database (DSLD) is created and managed by Office of Dietary Supplement and U.S. National Library of Medicine in the National Institutes of Health. Each product label includes:

- Product information (including product name, statement of identity, serving information, and target groups)
- Dietary supplement facts (including active ingredients, % daily value);
- Label statements (including formulation, precaution, and suggested use)
- Manufacturer information

A total of 65499 dietary supplement label information was downloaded using the API http://dsld.nlm.nih.gov/dsld/api/label/ with the ID of the label appended to it. The result was in JSON format which was stored using MongoDB on AWS. The label ID for all the available supplements was downloaded in CSV format from the DSLD website. All the processing was done on an AWS instance.

To extract the side effect terms in the supplement labels SIDER 4.1 (Side Effect Resource) database was used. It contains information on 1430 marketed drugs and their recorded side effects (a total of 5868 side effects). SIDER uses the MedDRA (Medical Dictionary for Regulatory Activities) dictionary to extract side effects from drug labels. Although the SIDER database does not include dietary supplements, the side effect terms used for both drugs and supplements are similar. Hence it was used to standardize the side effect terms contained in each supplement label.

Workflow



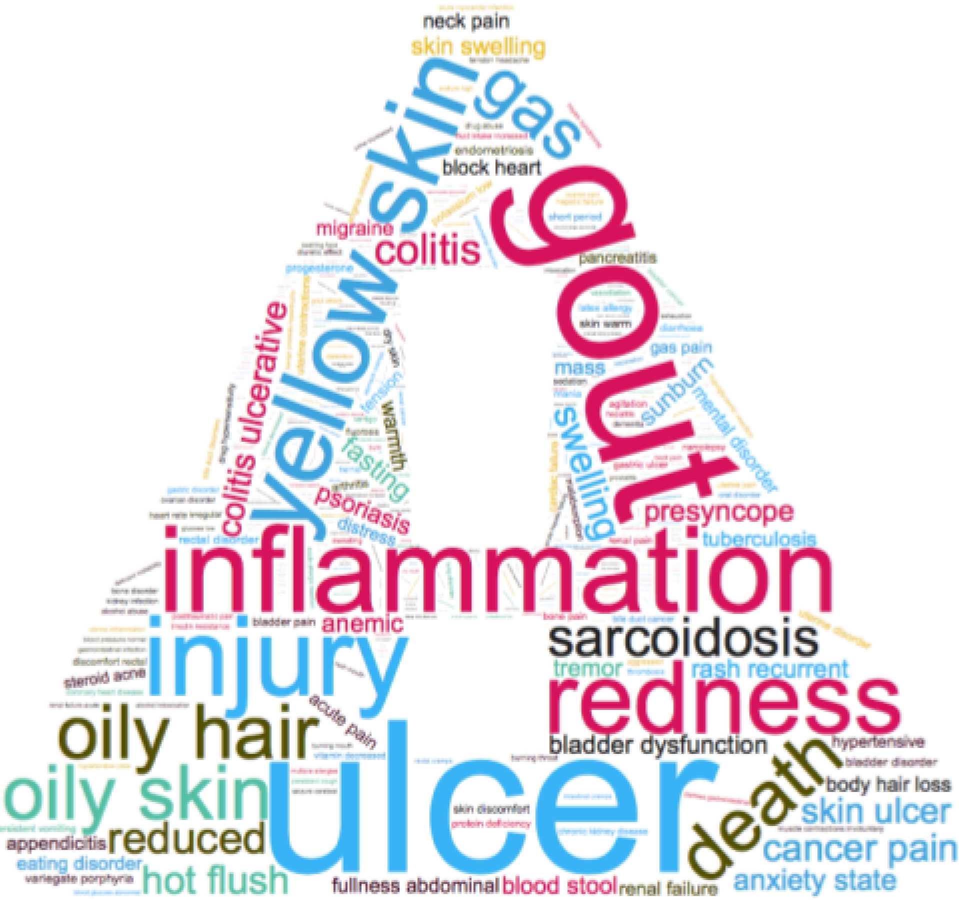
Tools Used

AWS, MongoDB, Gensim, NLTK, Scikit Learn, pyLDAvis

Document Preparation

- Extract Precaution/Warning statements from supplement labels

- Tokenize the text in the Precaution/Warning statements
  - Split into word tokens
  - Convert to lowercase
  - Remove punctuation from each token
  - Filter out tokens that are not alphabetic
  - Remove tokens of length < 3
  - Filter out tokens that are stop words
- Extract side-effect terms using SIDER database which contains 6123 unique side effects
- Join side effect terms which have more than one word with ‘_’
- Remove supplements which do not contain any side effect terms
- Left with ~16966 Dietary Supplement ‘documents’ with 662 unique side effect ‘tokens’



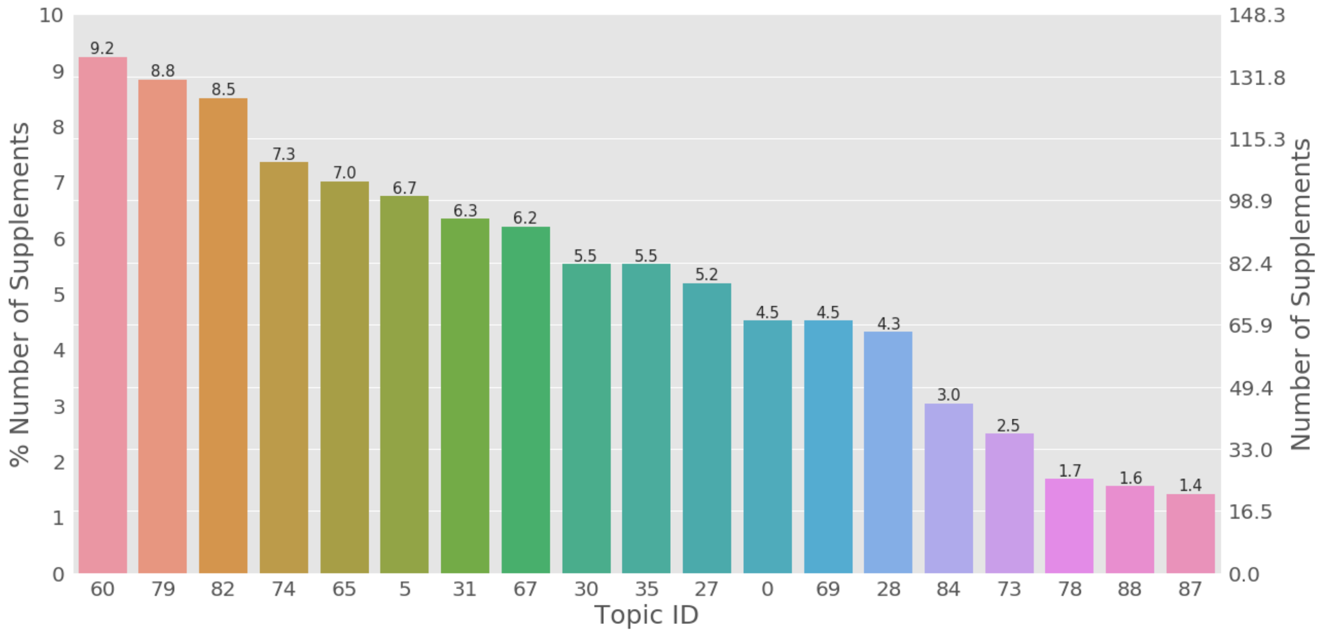
Word cloud of the 662 unique side effect ‘tokens’

Topic Modeling

- Implemented the Latent Dirichlet Allocation (LDA) model since it assumes that all the documents in a collection can be described by a group of topics. Python's gensim package was used to generate the LDA topics.
- Tested different number of topics ranging from 20 to 200 and looked at the topics for each. Finally, picked 90 topics since it had the best perplexity score and it gave the best topics.
- Topic terms with probability less than 0.02 were not considered for defining that topic
- Each supplement 'document' was assigned a topic which had the maximum conditional probability given the 'document'
- Grouped supplements which were assigned the same topic

Topic Selection

- Selected topics that have more than 2 terms with probability greater than 0.02
- Analyzed topics that have 10-150 supplements under it
- Left with 19 topics



Topic Evaluation

For each topic found the representative ingredient by doing a TF-IDF count of the supplement ingredients in that topic group. Then compared the topic side effect terms with the side effects of the representative ingredient obtained from WebMD website. The table below shows 4 of the topics obtained using LDA and their representative ingredient.

Representative Ingredient	Topic Terms	WebMD Side Effects
Topic 60: Caffeine	'difficulty_sleeping', 'anxiety', 'nervousness', 'palpitations', 'headache'	Causes insomnia, nervousness and restlessness, stomach irritation, nausea and vomiting, increased heart rate and respiration
Topic 82: Kava kava extract	'fatigue', 'vomiting', 'yellow_skin', 'injury', 'liver_injury', 'abdominal_injury', 'drowsiness'	Serious illness, including liver damage. Early symptoms of liver damage include yellowed eyes and skin (jaundice), fatigue, and dark urine.
Topic 67: Niacin	'redness', 'clotting', 'feeling_hot', 'skin_irritation', 'hot_flush', 'skin_bleeding'	common minor side effect is a flushing reaction. This might cause burning, tingling, itching, and redness of the face, arms, and chest, as well as headaches.
Topic 79: Calcium	'obstruction', 'intestinal_obstruction', 'pain', 'abdominal_pain', 'loose_stools', 'colon_obstruction', 'abdominal_disorder'	Calcium supplements cause few, if any, side effects. But side effects can sometimes occur, including gas, constipation and bloating. In general, calcium carbonate is the most constipating.

Conclusion

- Generated useful side-effect based grouping of dietary supplements without any prior knowledge
- Looking at the information of all the supplements within a group could provide insight for further studies

Future Work

- Enhance the technique to extract side-effects terms from supplement labels
- Extend this model to mine FDA drug labels to investigate potential interactions between dietary supplements and drugs

Reference

Wang Y, Gunashekar DR, Adam TJ, Zhang R. Mining Adverse Events of Dietary Supplements from Product Labels by Topic Modeling. *Studies in health technology and informatics*. 2017;245:614-618.