Predicting the effect of Genetic Variants to enable Personalized Medicine

Presented by CgA-*Team*

NYCDSA

Presentation Outline

1. Introduction
2. Workflow
3. EDA & Feature Engineering
4. NLP Model
5. Classification
6. Web scraping/Database
7. Summary

Introduction

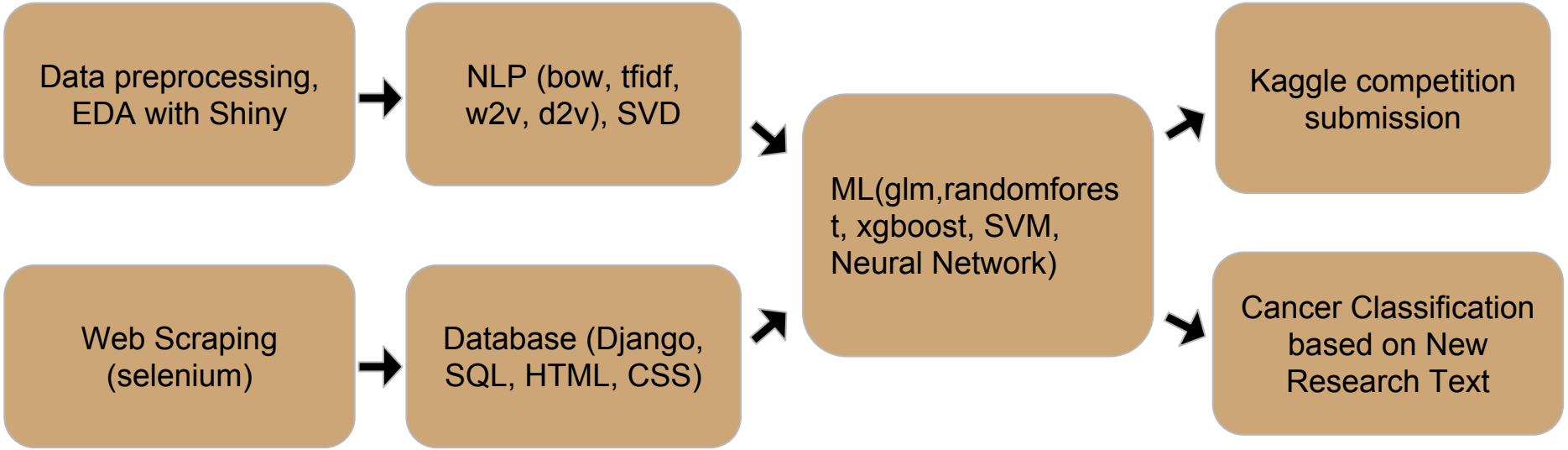
- Cancer tumors can have thousands of different genetic mutation variations
- Variations can be classified as contributors to tumor growth (drivers) or neutral mutations (passengers)
- Currently this classification is done manually
- Memorial Sloan Kettering Cancer Center opened a Kaggle Competition, asking participants to classify variations across nine mutually exclusive classes

Memorial Sloan Kettering Cancer Center

Data

- Training set - 3,321 variants
- Test set - 5,668 variants
- Variables
  - Gene
  - Variation
  - Text from academic papers used to classify the variant
  - Class (for training set)
- Missingness
  - 5 missing text values

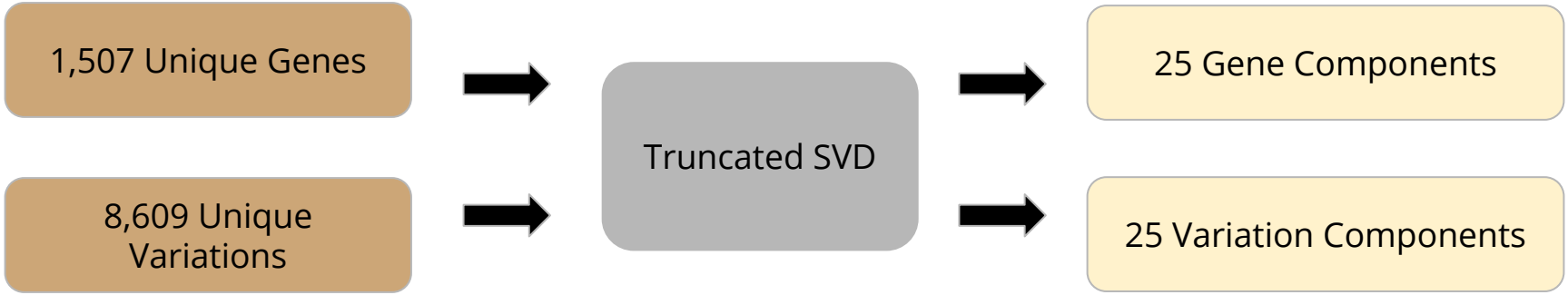
Project Workflow



EDA - R Shiny Demo

- See Shiny App

Truncated SVD



Number of components:

- 5 gene/10 variation performed better in multinomial logistic regression and random forest models without vectorized text features
- 25 gene/25 variation performed better in combination with vectorized text features

Word Embedding - Count based

- Count-based vectorization - does not preserve order, ignores semantics
  - Bag-of-words - essentially simple word count
  - TF-IDF - compares term frequency in each entry vs. entire corpus

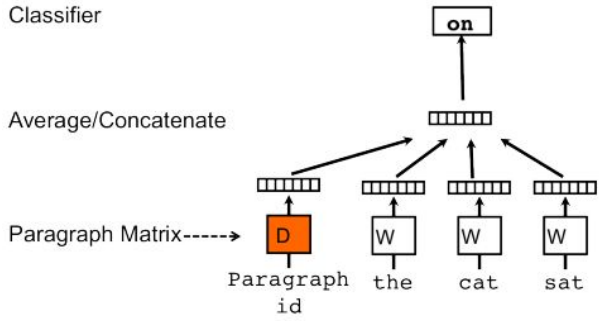
- tf-idf weight is simple the product of tf and idf weight.

$$W_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10}(N/df_t)$$

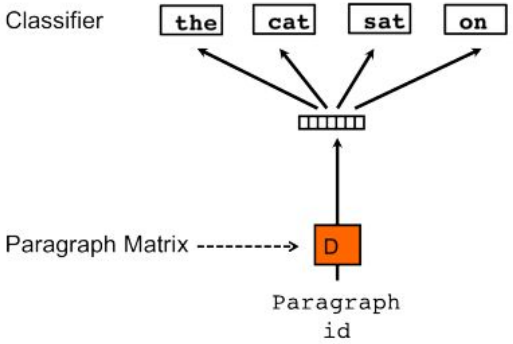
- Increases with number of occurrences within document.
- Increases with rarity of term in collection.

Word Embedding with Doc2Vec

PV-DM* (Distributed memory) - trains paragraph and word vectors together and averages them in the same space. The paragraph tags are treated as just another word token in the overall vectorization and help give words even more context.



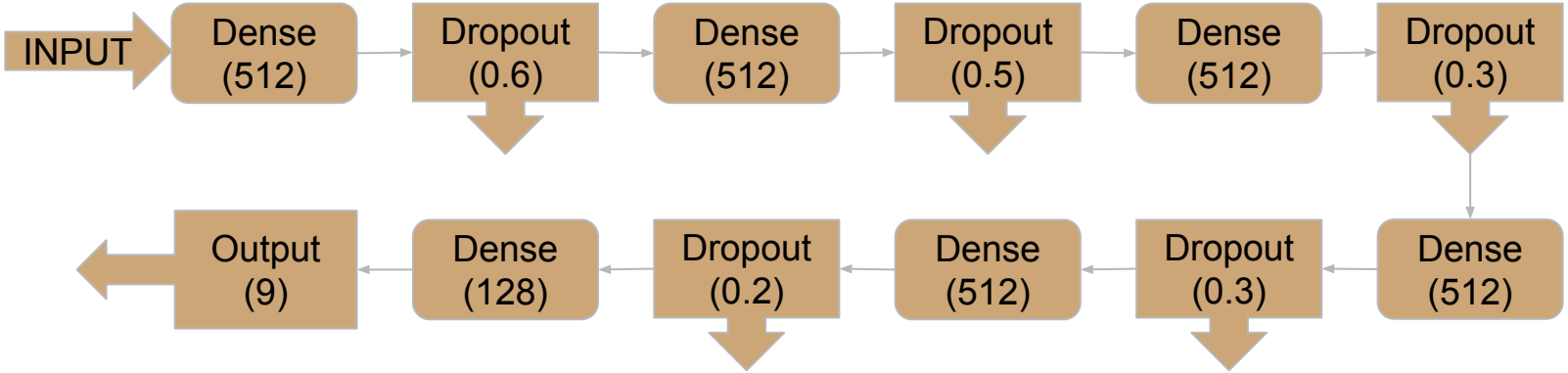
PV-DBOW (Distributed bag-of-words) - uses solely the paragraph vector to make inferences on context by sampling random words in the sliding window of each paragraph - can also train a skip-gram model alongside the paragraph vector.



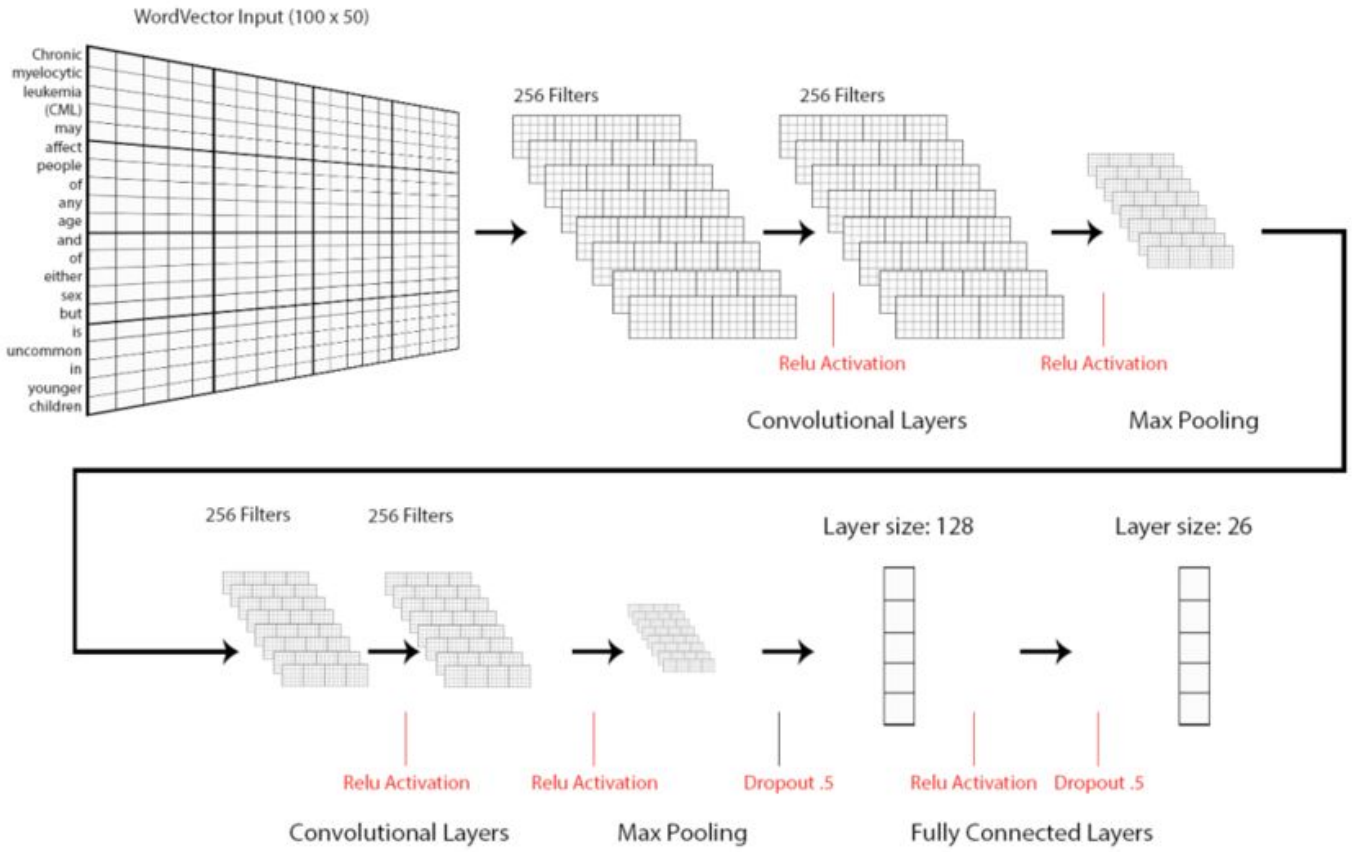
*default method employed by gensim

Deep Learning - ANN

Our first neural network consists of seven dense (fully-connected) layers with dropouts in between to prevent overfitting. All activation functions were rectified linear units (‘ReLU’).

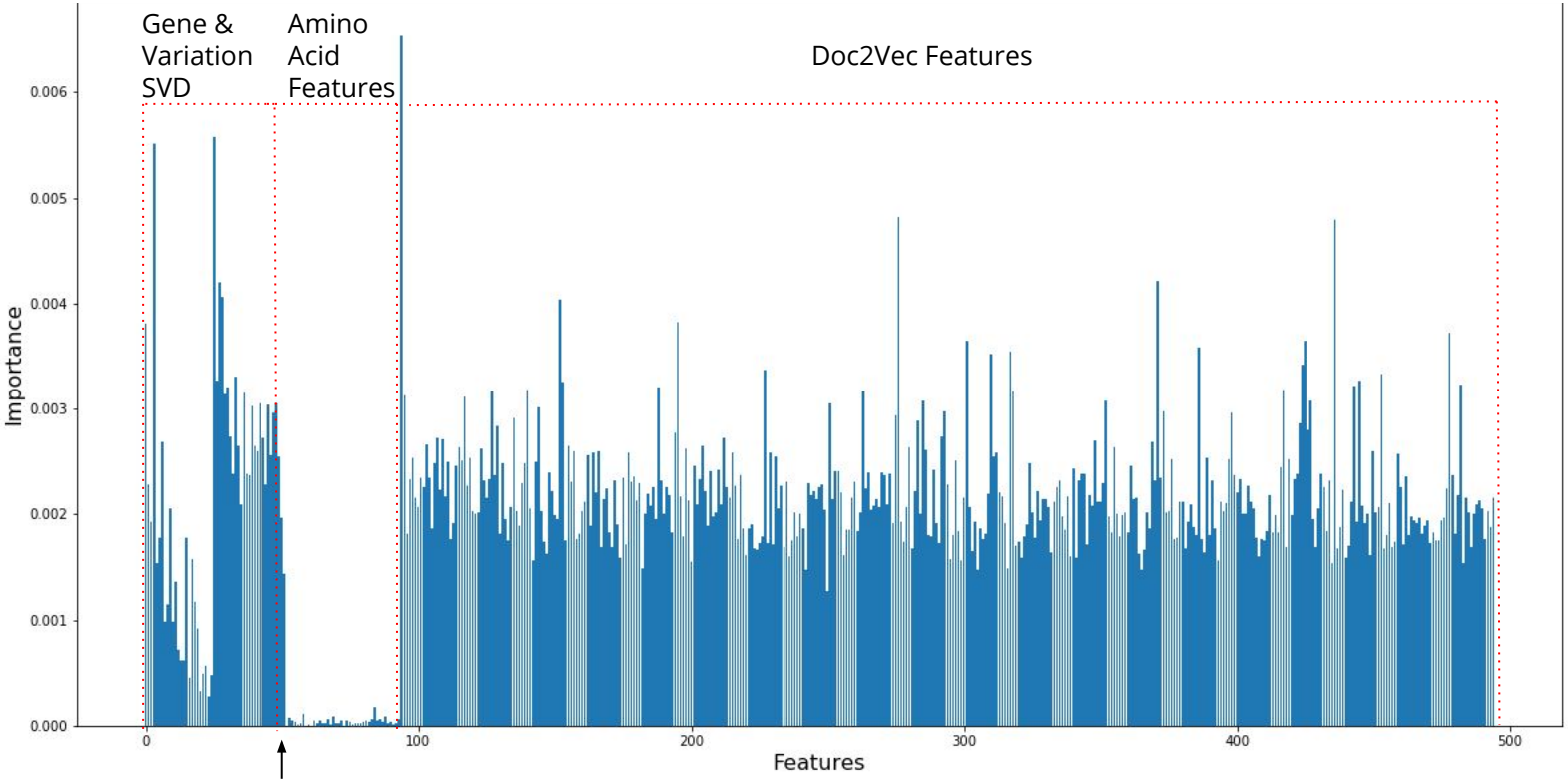


CNN



Hughes, Mark, et al. “Medical Text Classification Using Convolutional Neural Networks.”

Feature Importance (XGBoost)



Amino Acid Charge Difference: .002 importance
Hydrophobicity: .0015 importance

Machine Learning Models

ML Algorithm	Text Vectorization	Text Alterations	Important Hyperparameters	Accuracy	Notes
Support Vector Classifier	KATIE				
Random Forest	TFIDF	Stop words	Max depth =15, n_estimates = 100, CV	.569	Selected keywords context
Multinomial Logistic Regression	Doc2Vec	Stop words	C = .1, L1 CV	.595	Ok at predicting all classes
Multinomial Naive Bayes	Doc2Vec	Stop words	MinMaxScaler	.486	Overpredicts popular classes
Support Vector Classifier	Doc2Vec	Stop words		.609	Underpredicts popular classes
XGBoost	Doc2Vec	Stop words	Eta = .01, n_estimators = 1000, max_depth = 15	.701	Good at predicting all classes

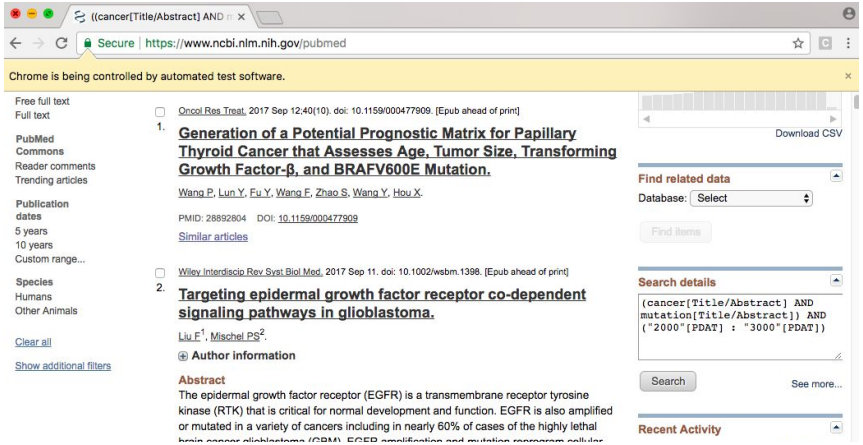
Machine Learning Models - Neural Networks

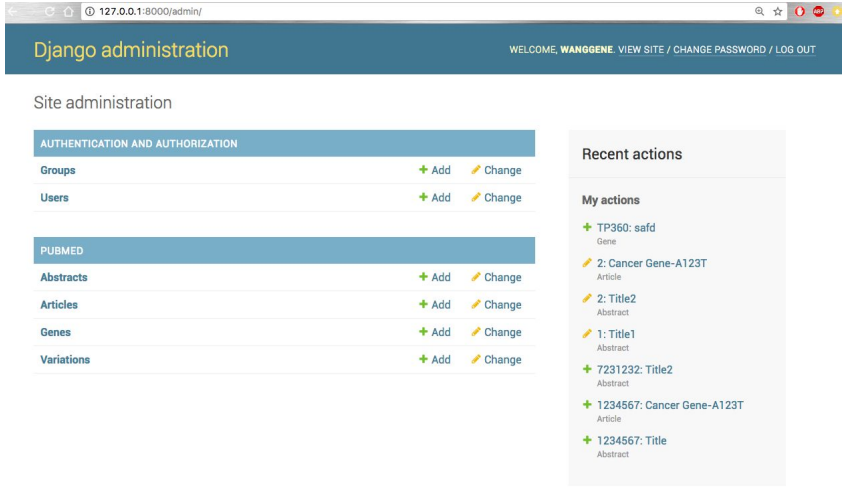
ML Algorithm	Text Vectorization	Text Alterations	Important Hyperparameters	Accuracy	Notes
Neural Network	TF-IDF	Context Alterations	RNN, LSTM layer, num_words = 2000	.687	Best neural network
Neural Network	Doc2Vec	Stop words	6 fully-connected layers	.632	Underpredicts popular classes
Neural Network	Doc2Vec	Stop words	2 convolutional layers 6 fully-connected layers	.601	Ok at predicting all classes

Model Takeaways

- Models confused (class 2 ↔ class 7) (class 1 ↔ class 4)
- Changing ‘class_weights = balanced’ reduced accuracy, but fixed overprediction of popular classes (important given our task at hand)
- XGBoost - most effective ML technique
- Doc2Vec - most effective text vectorization technique
- Ensemble of 8 models - performed worse than XGBoost model

Web Scraping and Database development





- NIH Pubmed: https://www.ncbi.nlm.nih.gov/pubmed/
- Key words: “Cancer”, “Gene”, “Mutation”, “SNP”
- Web scraping using Selenium.
- ~ 37000 articles

Cancer Research Text Classification & Recommendations

NYC DATA SCIENCE ACADEMY

Cancer Genetic Variants Classification *by CgA-Team*

Abstract Title List

- 28881380
  Missed therapeutic and prevention opportunities in women with BRCA-mutated epithelial ovarian cancer and their families due to low referral rates for genetic counseling and BRCA testing: A review of the literature.
- 28880737
  Cost-effectiveness of osimertinib in the UK for advanced EGFR-T790M non-small cell lung cancer.
- 28880088
  Brain accumulation of ponatinib and its active metabolite N-desmethyl ponatinib is limited by P-glycoprotein (P-GP/ABCB1) and breast cancer resistance protein (BCRP/ABCG2).
- 28880013
  Osimertinib (AZD9291) decreases programmed death ligand-1 in EGFR-mutated non-small cell lung cancer cells.
- 28879638
  Impact of Etoposide on BRCA1 Expression in Various Breast Cancer Cell Lines.
- 28879519
  Safety, tolerability, and pharmacokinetic profile of dabrafenib in Japanese patients with BRAF V600 mutation-positive solid tumors: a phase 1 study.
- 28879469

NYC DATA SCIENCE ACADEMY

Cancer Genetic Variants Classification *by CgA-Team*

Unraveling genetic predisposition to familial or early onset gastric cancer using germline whole-exome sequencing.

Pubmed ID: 28875981

Class: 8

Recognition of individuals with a genetic predisposition to gastric cancer (GC) enables preventive measures. However, the underlying cause of genetic susceptibility to gastric cancer remains largely unexplained. We performed germline whole-exome sequencing on leukocyte DNA of 54 patients from 53 families with genetically unexplained diffuse-type and intestinal-type GC to identify novel GC-predisposing candidate genes. As young age at diagnosis and familial clustering are hallmarks of genetic tumor susceptibility, we selected patients that were diagnosed below the age of 35, patients from families with two cases of GC at or below age 60 and patients from families with three GC cases at or below age 70. All included individuals were tested negative for germline CDH1 mutations before or during the study. Variants that were possibly deleterious according to in silico predictions were filtered using several independent approaches that were based on gene function and gene mutation burden in controls. Despite a rigorous search, no obvious candidate GC predisposition genes were identified. This negative result stresses the importance of future research studies in large, homogeneous cohorts.European Journal of Human Genetics advance online publication, 6 September 2017; doi:10.1038/ejhg.2017.138.

Keywords:

Contact | LinkedIn | Twitter | Google+

```
In [70]: search('Impact of Etoposide on BRCA1 Expression i
Out[70]: set()
```

```
In [72]: import pandas as pd
         doc2vec_df = doc2vec_model.docvecs.most_similar('
         pd.DataFrame(doc2vec_df, columns=['title', 'Dista
```

Out[72]:

	title	Distance
0	BRCA1 and FOXA1 proteins coregulate the expres...	0.786017
1	Mutations in BRCA2 and taxane resistance in pr...	0.769862
2	Breast cancer cell response to genistein is co...	0.766796
3	BRCA1-Mutated Estrogen Receptor-Positive Breas...	0.766661
4	BRCA1 Mutation Leads to Deregulated Ubc9 Level...	0.762232
5	Binding of CtIP to the BRCT repeats of BRCA1 i...	0.758850
6	BRCA2 is ubiquitinated in vivo and interacts w...	0.755799
7	The RING heterodimer BRCA1-BARD1 is a ubiquiti...	0.754895
8	A delayed chemically induced tumorigenesis in ...	0.751110
9	VEGFR3 inhibition chemosensitizes ovarian canc...	0.749491

```
In [ ]:
```

Summary

- Went a step further than Kaggle to create an app that can be utilized by oncologists to streamline classification
- All results are irrelevant due to leaked data
- New data will be posted in ~2 weeks, other valid models at the top of the leaderboard (read: overfit) will likely perform worse on new data

Summary .2

- EDA and Shiny app development
- Feature engineering, including AA info and keyword context
- Text preprocessing, NLP and ML, Deep Learning with RNN, CNN
- Our best submission had a multiclass loss of .65428, which ranks 407 out of 1142
- Cancer classification with new research text input

Future work

- Use of only Doc2Vec to find similarities between documents to see which ones lead to misclassification
- Model stacking / further ensembling
- Keep updating database, optimize the UI

Thank you!

Presented by CgA-*Team*