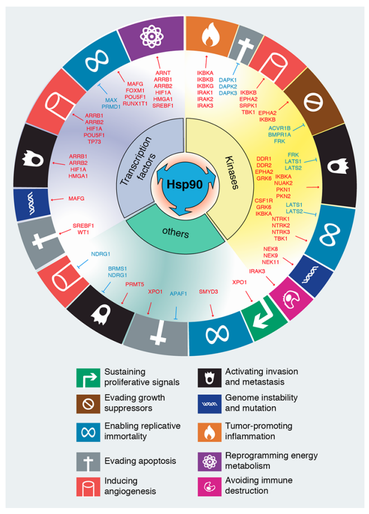
Using NLP with Machine Learning to classify Cancer Gene Mutations

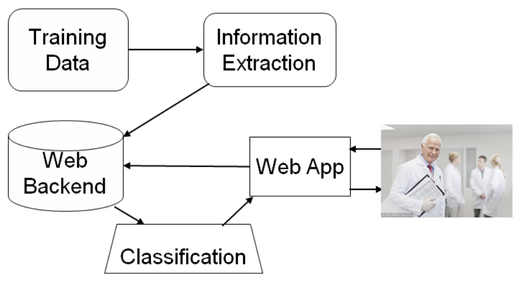
Trello: https://trello.com/b/tEEyg11e/cancerdetection Github: https://github.com/ShuangZhao95/EC601-Cancer-Detection
Group Members: Shuang Zhao, Lijun Xiao, Zhexi Zhang, John Curci

Introduction

A cancer tumor can have thousands of genetic mutations. It is very time-consuming for clinical pathologists to review and classify every single genetic mutation based on evidence from text-based clinical literature manually.



	ID	Gene	Variation	Text
0	0	FAM58A	Truncating Mutations	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	CBL	W802*	Abstract Background Non-small cell lung canc...
2	2	CBL	Q249E	Abstract Background Non-small cell lung canc...
3	3	CBL	N454D	Recent evidence has demonstrated that acquired...
4	4	CBL	L399V	Oncogenic mutations in the monomeric Casitas B...
5	5	CBL	V391I	Oncogenic mutations in the monomeric Casitas B...
6	6	CBL	V430M	Oncogenic mutations in the monomeric Casitas B...
7	7	CBL	Deletion	CBL is a negative regulator of activated recep...
8	8	CBL	Y371H	Abstract Juvenile myelomonocytic leukemia (JM...
9	9	CBL	C384R	Abstract Juvenile myelomonocytic leukemia (JM...



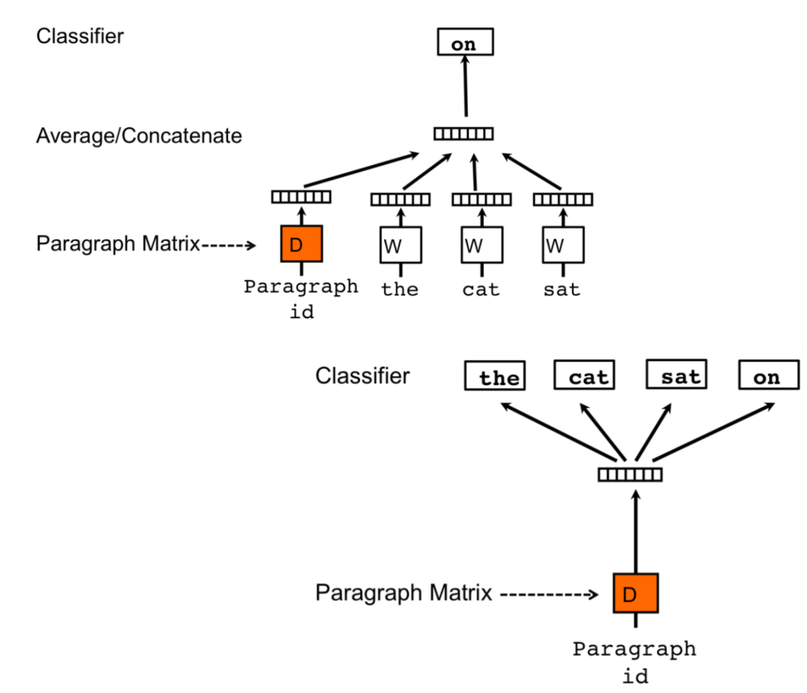
High level system diagram

Our project is designed for scientists who would like to extract information from large number of scientific papers quickly without reading them and get the classification of the cancer gene mutations.

The data set we got is from Kaggle and it contains thousands of different gene mutations along with a long description for each of them.

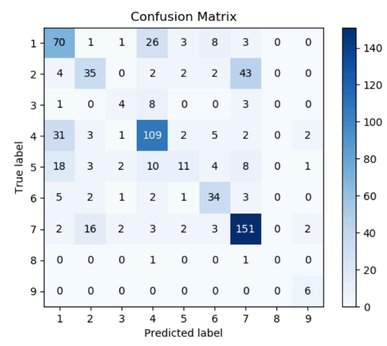
Processing Data: Doc2Vec

- Preprocessing: Pandas, Regex to remove symbols, stopwords.
- Based on Word2Vec.
- Words are mapped to unique vectors.
- Have a unique paragraph vector for each paragraph.
- Is essentially plug-and-play with Python.
- Is the only widely used document embedding model.

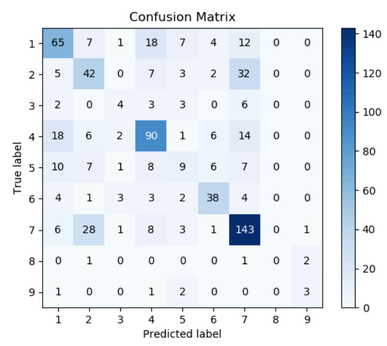


Model

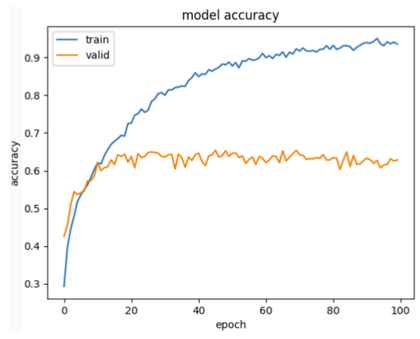
- Used neural network model with two hidden layers
- ReLU activation function – kills negative layer connections
- Implemented with Keras using Tensorflow backend
- Output layer creates nine class probability vector



Primary Model with Doc2Vec: 63%

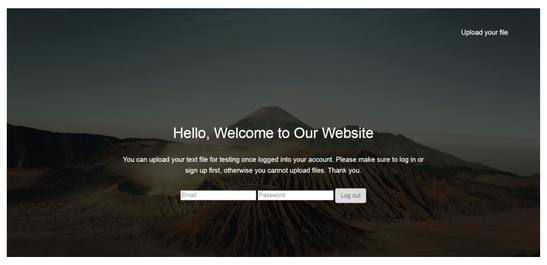


LSTM-based Model: 59%



Website

- Frontend: We use HTML to write the webpage.
- Backend: Flask is a microframework for Python.
- Deploy:
  - Firebase: Firebase authentication
  - AWS cloud machine



User Guide

[1] Sign up with your email address and create your password.
[2] Once logged in, you will see a upload file button on upper right corner.
[3] Upload your file contains the description of the gene. The file could be in the following format.

```
ID,Text
1|| Abstract The Large Tumor Suppressor 1 (LATS1) is a serine/threonine kinase and tumor
recently been identified as a central player of the emerging Hippo signaling pathway, whi
stem cell differentiation and renewal, etc. Although mounting evidence supports a role of
at the molecular level is not fully understood. Recently several positive regulators of L
negatively regulated is still largely unknown. We have recently identified Itch, a member
regulator of LATS1. However, whether other ubiquitin ligases modulate LATS1 stability and
family using over-expression and short-interference RNA knockdown approaches, we have ide
We have provided in vitro and in vivo evidence that WWP1 is essential for LATS1 stability
polyubiquitination and the 26S proteasome pathway. Importantly, we also showed that degra
proliferation in breast cancer cells. Since WWP1 is an oncogene and LATS1 is a tumor supp
```

BOSTON UNIVERSITY

Visit our Website: https://coolproject-visit.firebaseapp.com