



# DupQU - Duplicate Question Pair Prediction using NLP

## Guided By:

Radhika Mamidi

## Team Details:

Awani Rawat (20172029)

Shivani Sethi (20172024)

Surbhi Goyal (20172023)

Shubham Verma (20172035)

## Problem Statement

To predict which of the provided pairs of questions contain two questions with the same meaning





## Why??

- Redundant Questions
- More time to seek for best answer
- More time to write different answers for questions with same intent
- High Quality answers
- Useful in QA forums like Quora, etc.



## Domain Background

- Quora is a place to gain and share knowledge about anything.
- Over 100 million people visit Quora every month.
- Thus, asking multiple questions is wholly possible.
- It causes **seekers** to take more time to find best answers and **writers** to answer multiple versions of question with same intent.
- Canonical questions (~unique) are thus preferred in these forums.



## Dataset

- **Opensource:** Provided by Quora.
- **id:** the id of a training set question pair.
- **qid1, qid2:** unique ids of each question (only available in train.csv).
- **question1, question2:** the full text of each question.
- **is\_duplicate:** the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.



## Approach

- Problem type: Binary classification
- Algorithm type to use: Supervised learning

### Steps:

- Preprocessing of data
- Feature Extraction: tf-idf, wc, etc.
- Training Model: SVM, RF, K-nn, etc.
- Model Evaluation and Validation



## TIMELINE

- [Feb, 19] Requirements Study: Study about the techniques required for the project viz. ML Models, advanced NLP Techniques, semantic similarity, etc.
- [Mar, 19] Baseline Model: Training of model like SVM to get a start using features like tf-idf, word count, etc. and record the results.
- [Apr, 19] Final Model: Thinking of other models to use like Logistic Regression, RF model, KNN, etc. and retain the best one.

A decorative graphic on the left side of the slide. It consists of a blue parallelogram and a light green parallelogram, both tilted at an angle. The blue shape is in the foreground, and the green shape is partially behind it. They are set against a dark blue background with subtle diagonal lines.

# **BASELINE MODEL**

## **- SVM**

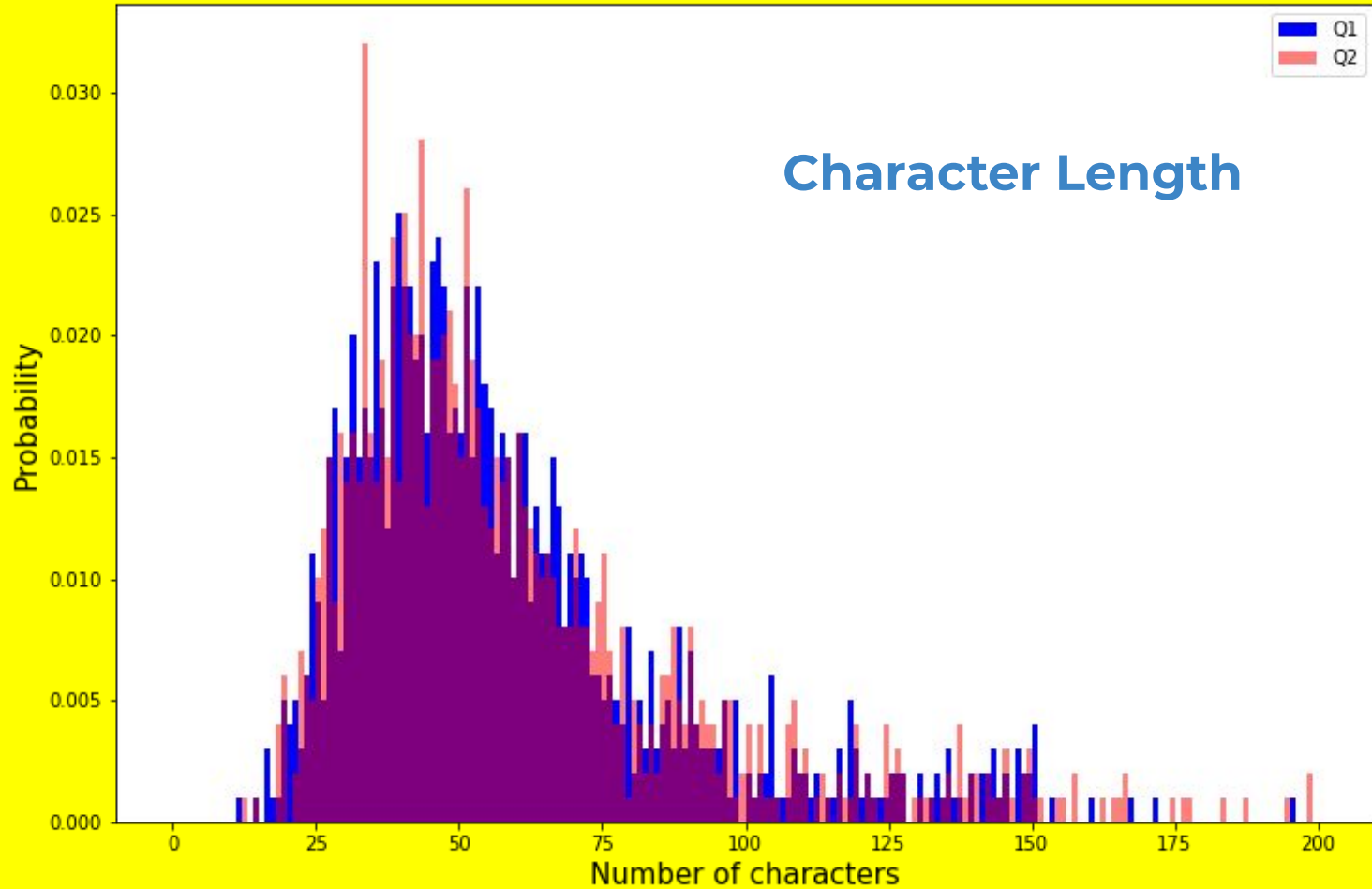


## FEATURES USED

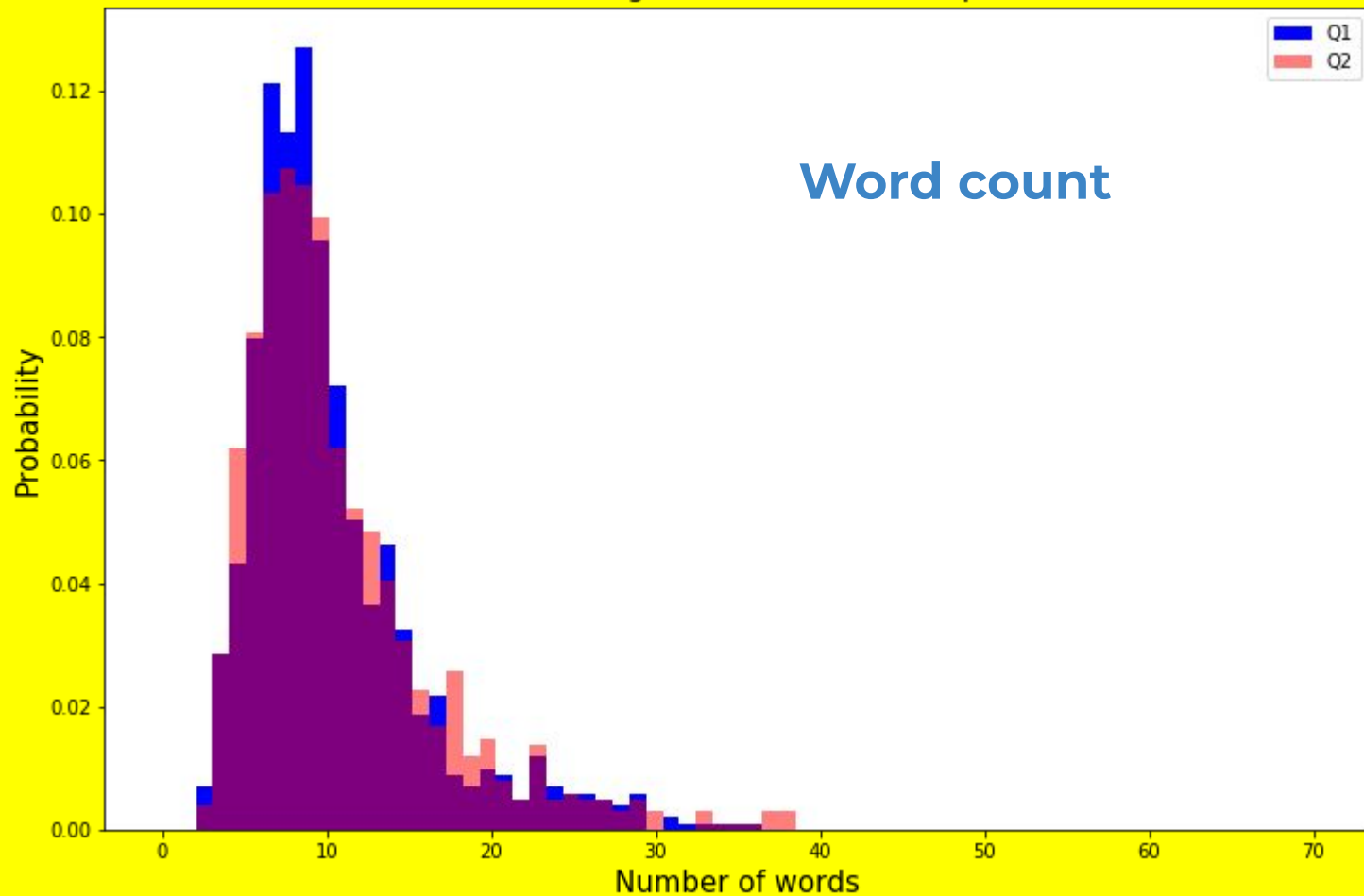
- Character length
- Word count
- Word share
- TF-IDF share

|   | q1chrln | q2chrln | q1_nword | q2_nword | word_share | TFIDF_share |
|---|---------|---------|----------|----------|------------|-------------|
| 0 | 66      | 57      | 14       | 12       | 0.434783   | 0.500000    |
| 1 | 51      | 88      | 8        | 13       | 0.200000   | 0.265928    |
| 2 | 73      | 59      | 14       | 10       | 0.166667   | 0.222361    |
| 3 | 50      | 65      | 11       | 9        | 0.000000   | 0.000000    |
| 4 | 76      | 39      | 13       | 7        | 0.100000   | 0.285535    |

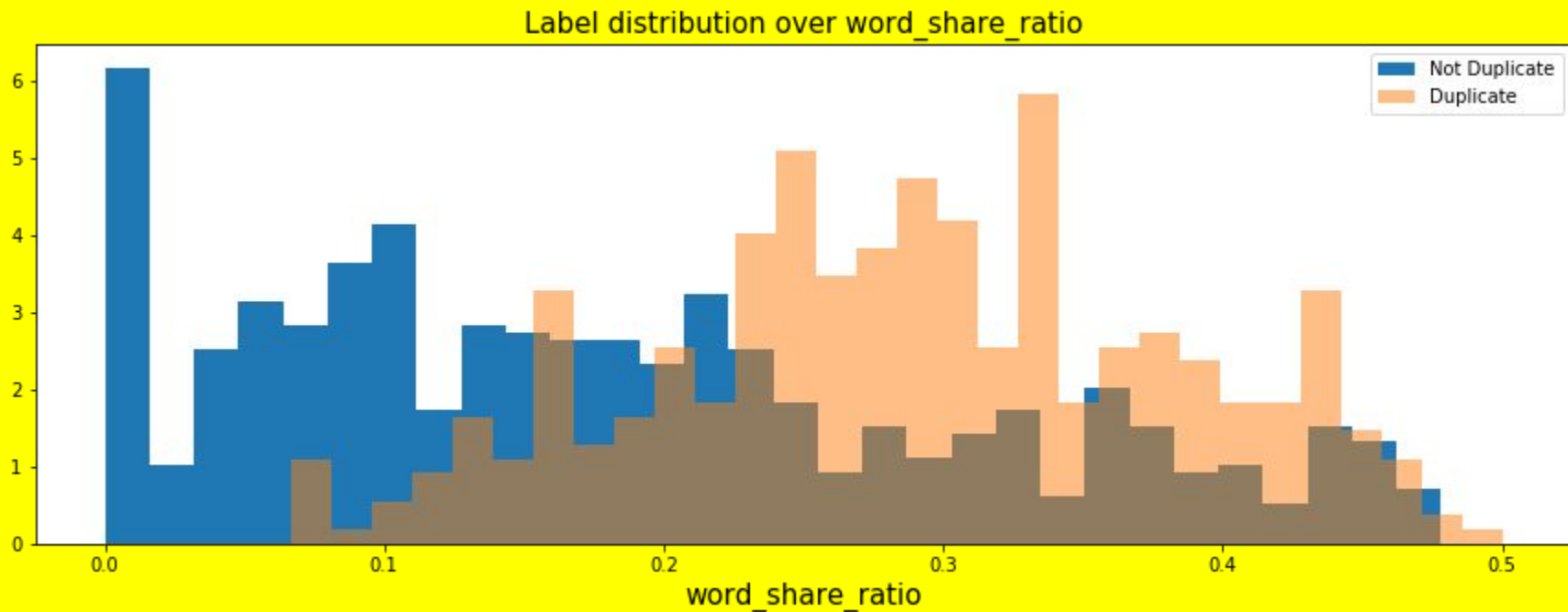
Normalised histogram of character count in questions



Normalized histogram of word count in questions



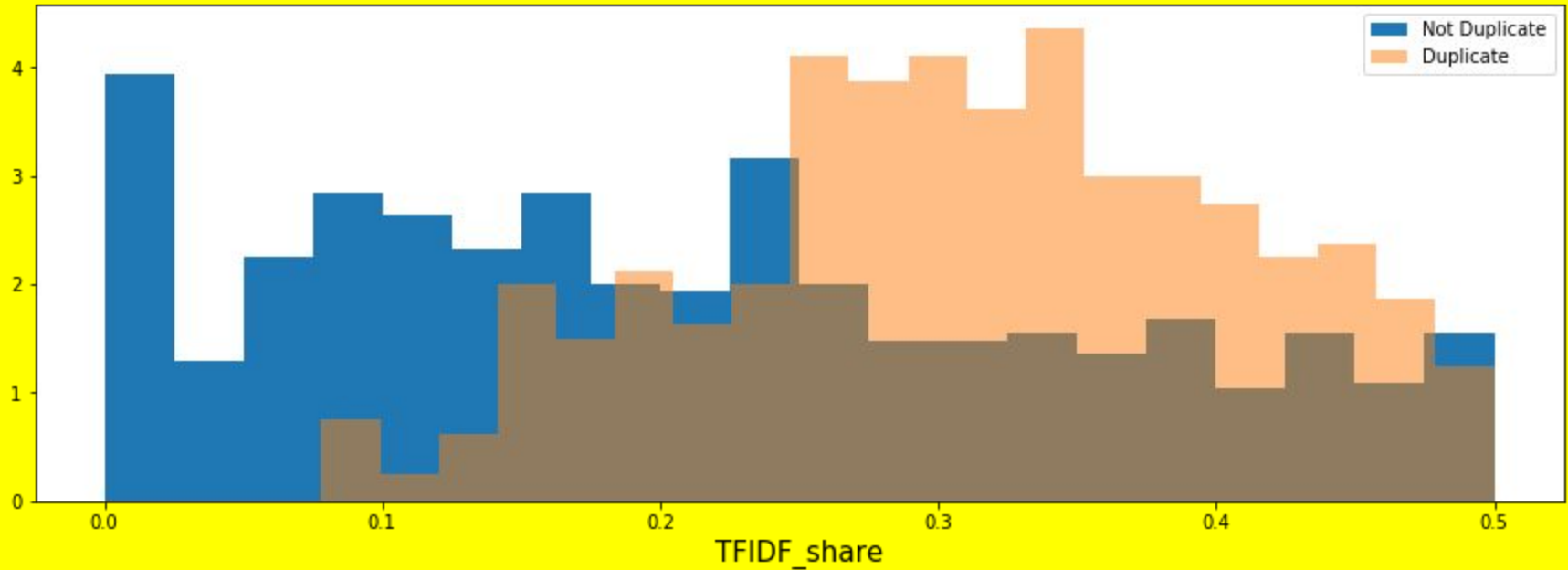
## Word share



**Word Share:**

$$= \text{len}(q1 \& q2) / (\text{len}(q1) + \text{len}(q2))$$

## TFIDF share

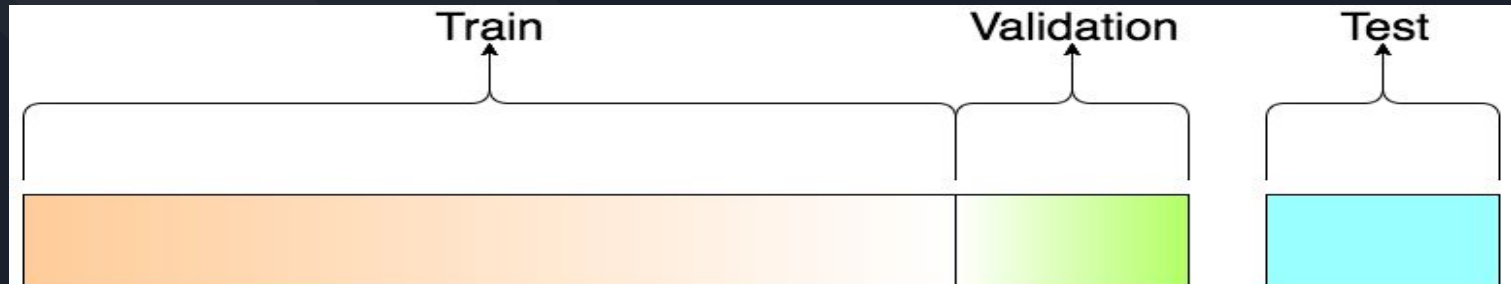


**TFIDF Share:**

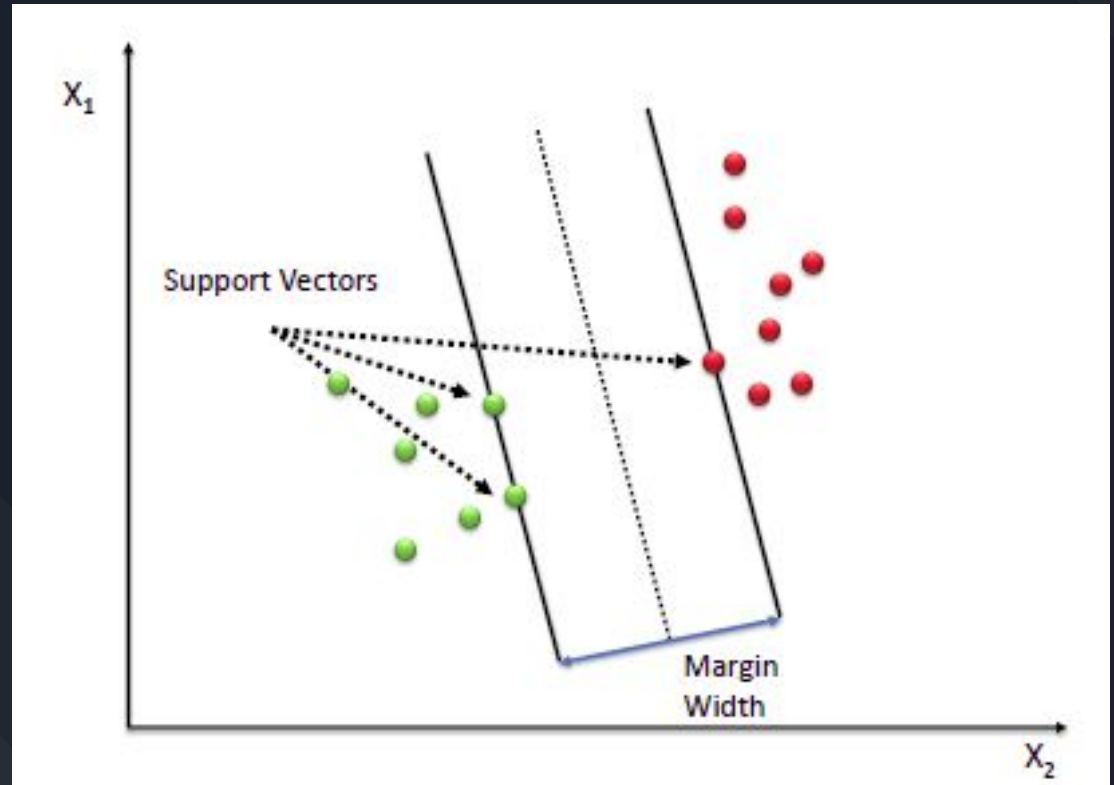
$$= \text{shared\_weight} / (\text{weight}(q1) + \text{weight}(q1))$$

## DIVIDING DATASET

- Training data (80 %)
  - Training (80%)
  - Validation (20%)
  - [Grid Search cross validation used]
- Testing data (20%)



# MODEL - SVM



# BASELINE MODEL RESULTS [60.5% accuracy]

| test_id | Q1  | Q2  | is_duplicate (in %) |
|---------|---|---|---------------------|
| 0       | How does the Surface Pro himself 4 con    | Why did Microsoft choose core m3 and not      | 26.08239            |
| 1       | Should I have a hair transplant at age 24 | How much cost does hair transplant requir     | 44.21242            |
| 2       | What but is the best way to send money    | What you send money to China?                 | 52.27987            |
| 3       | Which food not emulsifiers?               | What foods fibre?                             | 27.58848            |
| 4       | How "aberystwyth" start reading?          | How their can I start reading?                | 66.66284            |
| 5       | How are the two wheeler insurance fro     | I admire I am considering of buying insurar   | 34.74847            |
| 6       | How can I reduce my belly fat through a   | How can I reduce my lower belly fat in one    | 67.52494            |
| 7       | By scrapping the 500 and 1000 rupee no    | How will the recent move to declare 500 ar    | 41.01506            |
| 8       | What are the how best books of all time   | What are some of the military history book    | 72.79038            |
| 9       | After 12th years old boy and I had sex w  | Can a 14 old guy date a 12 year old girl?     | 28.6881             |
| 10      | What is the best slideshow app for And    | What are the best app for android?            | 76.88791            |
| 11      | What services are from Google: Facebo     | What social network (like Google, Faceboo     | 19.89593            |
| 12      | What if a cricket hits a batsman's he     | Should carbonated red balls and 8 yellow b    | 25.42344            |
| 13      | Just how do you learn fruity loops?       | How do Fruity Wrappers work?                  | 61.14768            |
| 14      | Why does Batman get kill in Batman v S    | In Batman v Superman, why reduce Lex Lut      | 50.90523            |
| 15      | When can I buy a SpaceX stock?            | Should I sell or buy LNKD stock?              | 55.39861            |
| 16      | Is it gouging and price fixing?           | What's the difference between intel of sor    | 28.68465            |
| 17      | Can a vacuum cleaner concentrate suck     | Could a vacuum cleaner suck get your eye c    | 58.70589            |
| 18      | I am 20 years old and I still a problem w | I am 20 years old and still have acne. It see | 48.01865            |
| 19      | What is it ai living in the middle class? | Why middle class?                             | 47.24792            |
| 20      | How matter at MIT? Will performing po     | I have passed 5 AP tests with scores trump    | 19.58628            |



A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is light green. They are positioned diagonally, with the blue one partially covering the green one.

# Feature Engineering

- Adding semantic features



**FUZZY FEATURES**

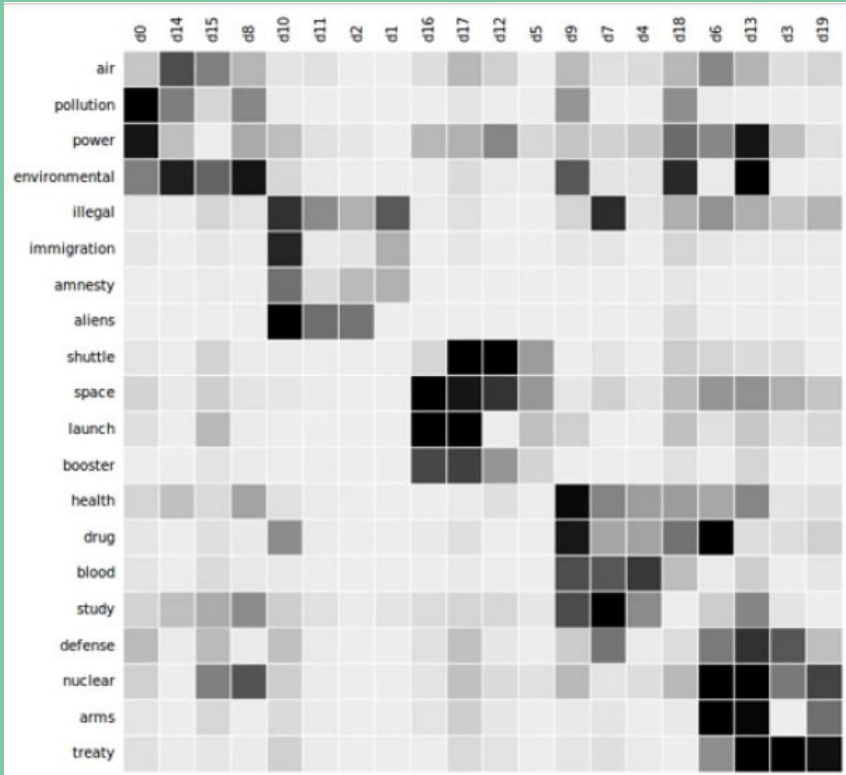
## ● Levenshtein distance

- It is a string metric for measuring the difference between two sequences.
- Informally, between two words it is the minimum number of single-character edits ([insertions](#), [deletions](#) or [substitutions](#)) required to change one word into the other. [[Edit distance](#)]

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

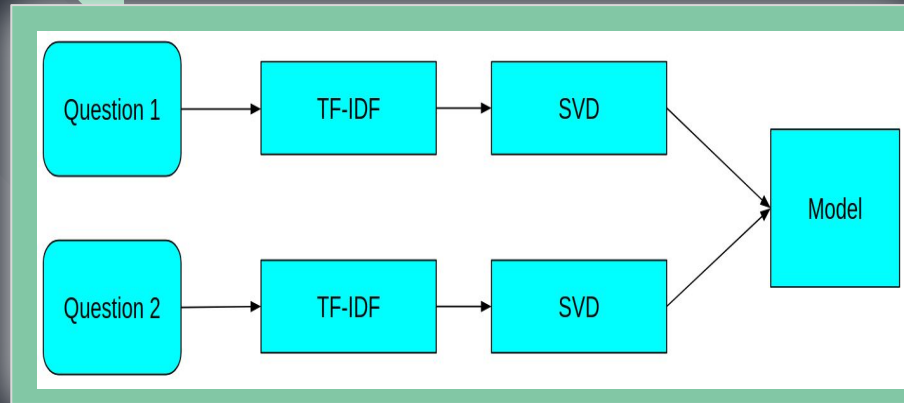
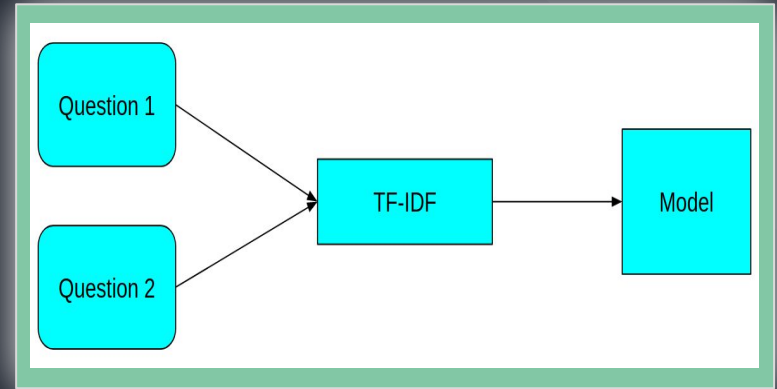
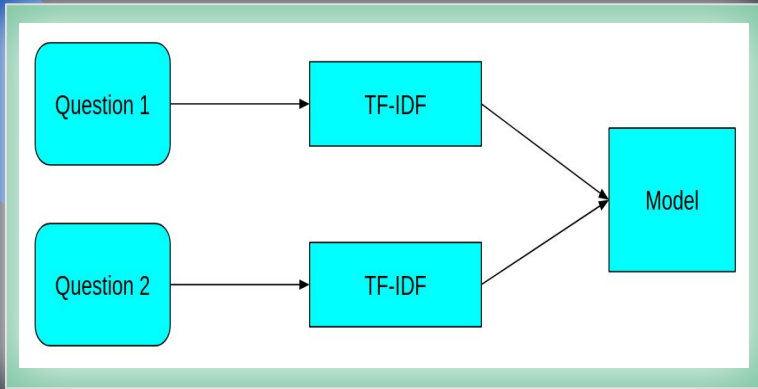
A blue parallelogram and a light green parallelogram are positioned on the left side of the slide, overlapping each other and the dark background.

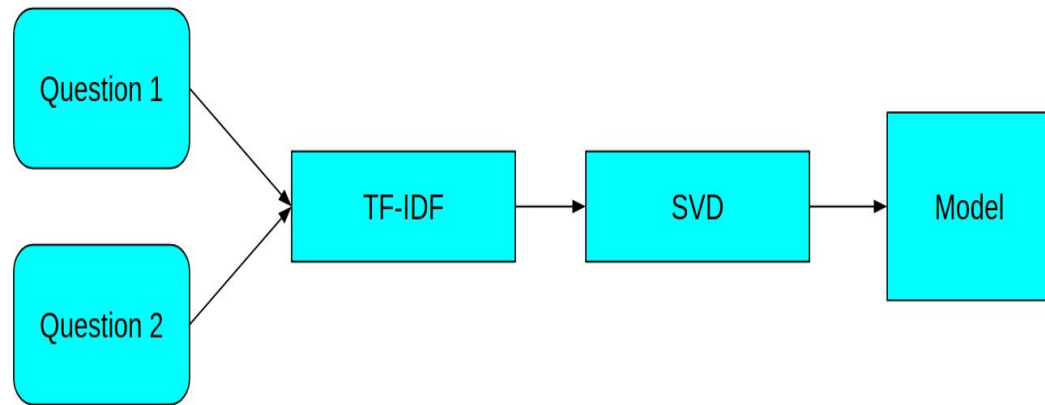
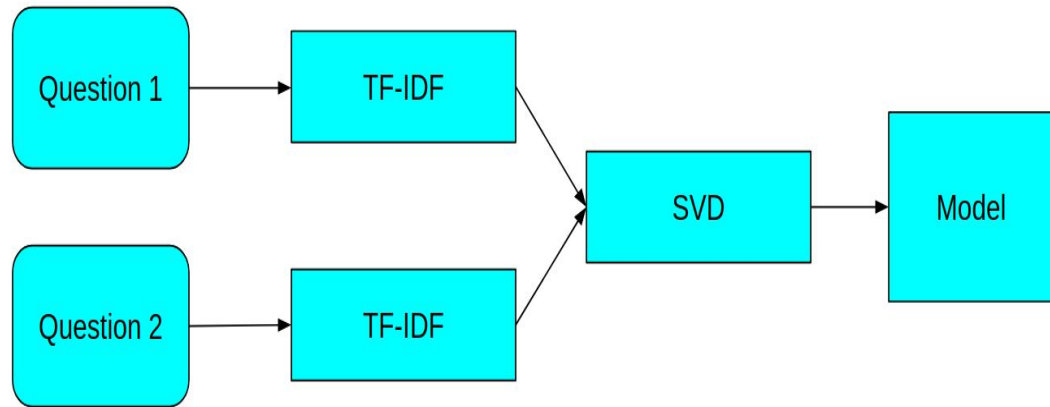
## **LSA (Latent Semantic Analysis) & SVD**



- LSA assumes that words that are close in meaning will occur in **similar pieces of text**.
- Uses SVD to handle sparsity of occurrence matrix (term-frequencies in documents)
- Various combinations of the features are taken into consideration to capture the similarity as maximally as possible

(to be contd . . .)



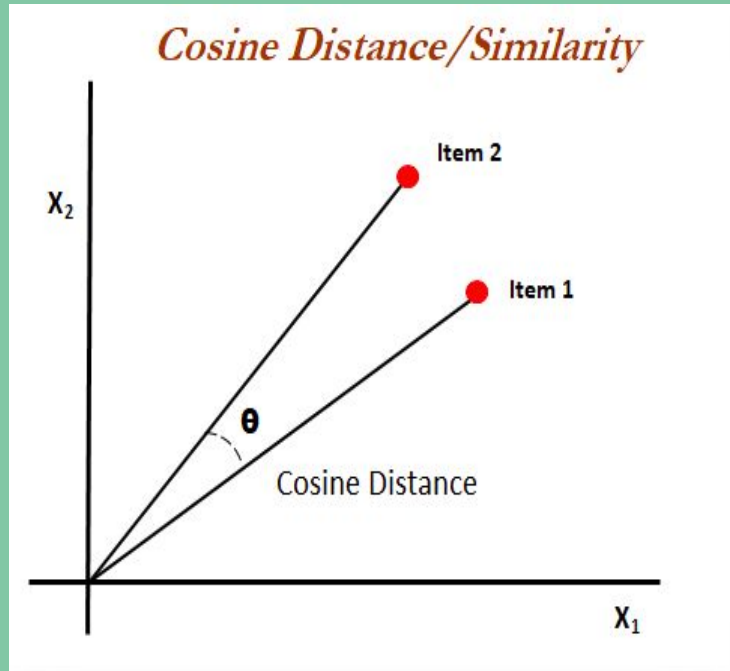


A blue parallelogram and a light green parallelogram are positioned on the left side of the slide, overlapping each other and the dark background.

## **Word2Vec features**



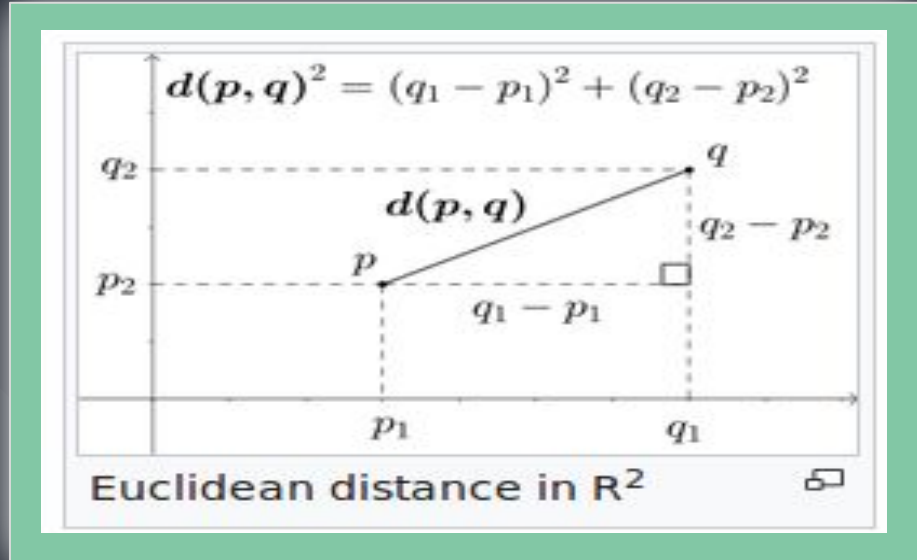
## • Cosine Distance/Similarity



- It is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.
- A soft cosine or ("soft" similarity) between two vectors considers similarities between pairs of features.
- It is a metric used to measure how similar the documents are irrespective of their size.

## • Euclidean Distance

- The Euclidean distance /metric is the "ordinary" straight-line distance between two points in Euclidean space.



## • Cityblock distance

- It is also known as Manhattan distance, boxcar distance, absolute value distance.
- It represents the **absolute differences** between coordinates of a pair of objects.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|,$$

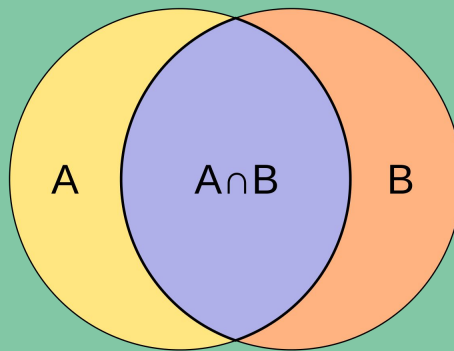
where  $(\mathbf{p}, \mathbf{q})$  are **vectors**

$$\mathbf{p} = (p_1, p_2, \dots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \dots, q_n)$$

## • Jaccard

- The Jaccard coefficient measures similarity between finite sample sets.
- It is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

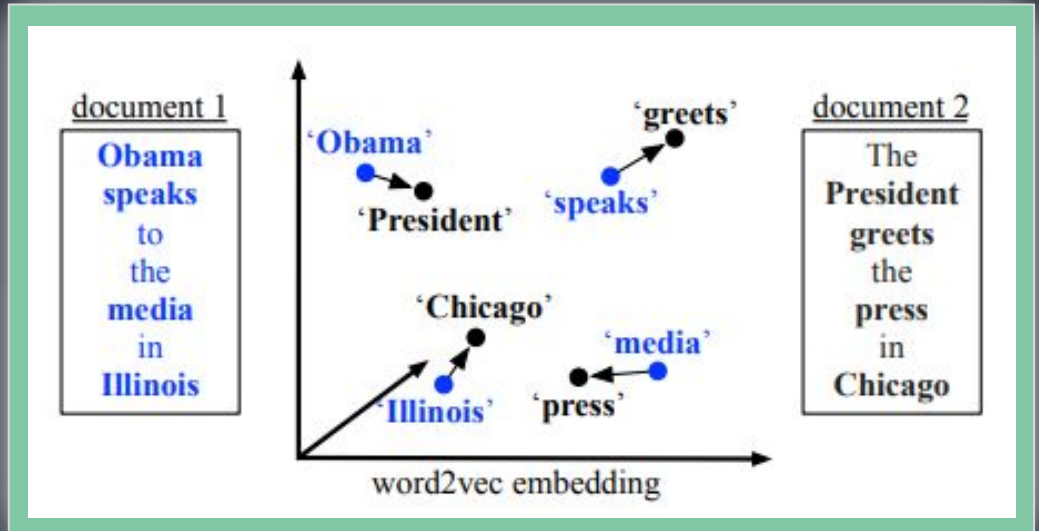


- **WMD (Word Mover Distance)** - Solution to overcome synonym problem
  - Chooses minimum transportation cost to transport every word from sentence 1 to sentence 2

$$\begin{aligned} \min_{\mathbf{T} \geq 0} \quad & \sum_{i,j=1}^n \mathbf{T}_{ij} c(i,j) \\ \text{subject to:} \quad & \sum_{j=1}^n \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \dots, n\} \\ & \sum_{i=1}^n \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \dots, n\}. \end{aligned}$$

## • WMD (Contd...)

- Sentence 1: Obama speaks to the media in Illinois
- Sentence 2: The president greets the press in Chicago
- Except the stop words, there is no common words among two sentences but both of them are talking about same topic (at that time).





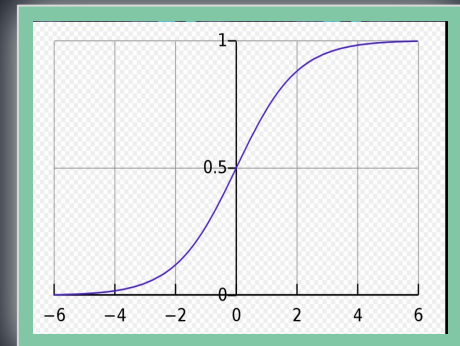
# **LOGISTIC REGRESSION**

- XGBoosting**

# • LOGISTIC REGRESSION

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- It is sensitive to the scale of the features.

The standard logistic function:







# ● XGBoosting

- The xgboost is a scalable, portable, and distributed gradient boosting library (a tree ensemble machine learning algorithm).
- It is the sag solver which requires a linear computational time in respect to the size of the data
- Being a gradient boosting algorithm, this learning algorithm has more variance (ability to fit complex predictive functions, but also to overfit) than a simple logistic regression afflicted by greater bias .

# • Results - 80% acc.

|    | Q1  | Q2  | prediction(in %) |
|----|---|---|------------------|
| 0  | Why do Swiss despise Asians?                      | Why do technical employees despise sales peopl... | 27.03            |
| 1  | What is the best/most memorable thing you've e... | What is the most delicious dish you've ever ea... | 64.08            |
| 2  | What are the types of immunity?                   | What are the different types of immunity in ou... | 65.12            |
| 3  | What is the quickest way to increase Instagram... | How can we increase our number of Instagram fo... | 29.77            |
| 4  | What universities does Rexnord recruit new gra... | What universities does B&G Foods recruit new g... | 33.86            |
| 5  | What is the stall speed and AOA of an f-14 wit... | Why did aircraft stop using variable-sweep win... | 21.77            |
| 6  | What are the questions should not ask on Quora?   | Which question should I ask on Quora?             | 45.3             |
| 7  | When will the BJP government strip all the Mus... | Why India does not apply the "Burma-Rohingya m... | 24.61            |
| 8  | Method to find separation of slits using fresn... | What are some of the things technicians can te... | 20.62            |
| 9  | What is the step by step guide to invest in sh... | What is the step by step guide to invest in sh... | 64.1             |
| 10 | Why are so many Quora users posting questions ... | Why do people ask Quora questions which can be... | 54.4             |
| 11 | What's one thing you would like to do better?     | What's one thing you do despite knowing better?   | 72.46            |


## ● Comparison:

SVM:

- *Features used:*
  - Basic statistical features
- *Accuracy* : 60.5 %

LR:

- *Features used:*
  - Fuzzy features (Levenshtein)
  - Semantic (LSA)
  - W2V features
- *Accuracy* : 50 %
- *Accuracy (XGBoosting)*: 80 %



**Project is live (with exe also):**

<https://github.com/vshubham8/DupQu>





Thank  
You