Questions Similarities (Quora Challenge)

Abdelhak NEZZARI

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Quora Challenge: Question Similarities

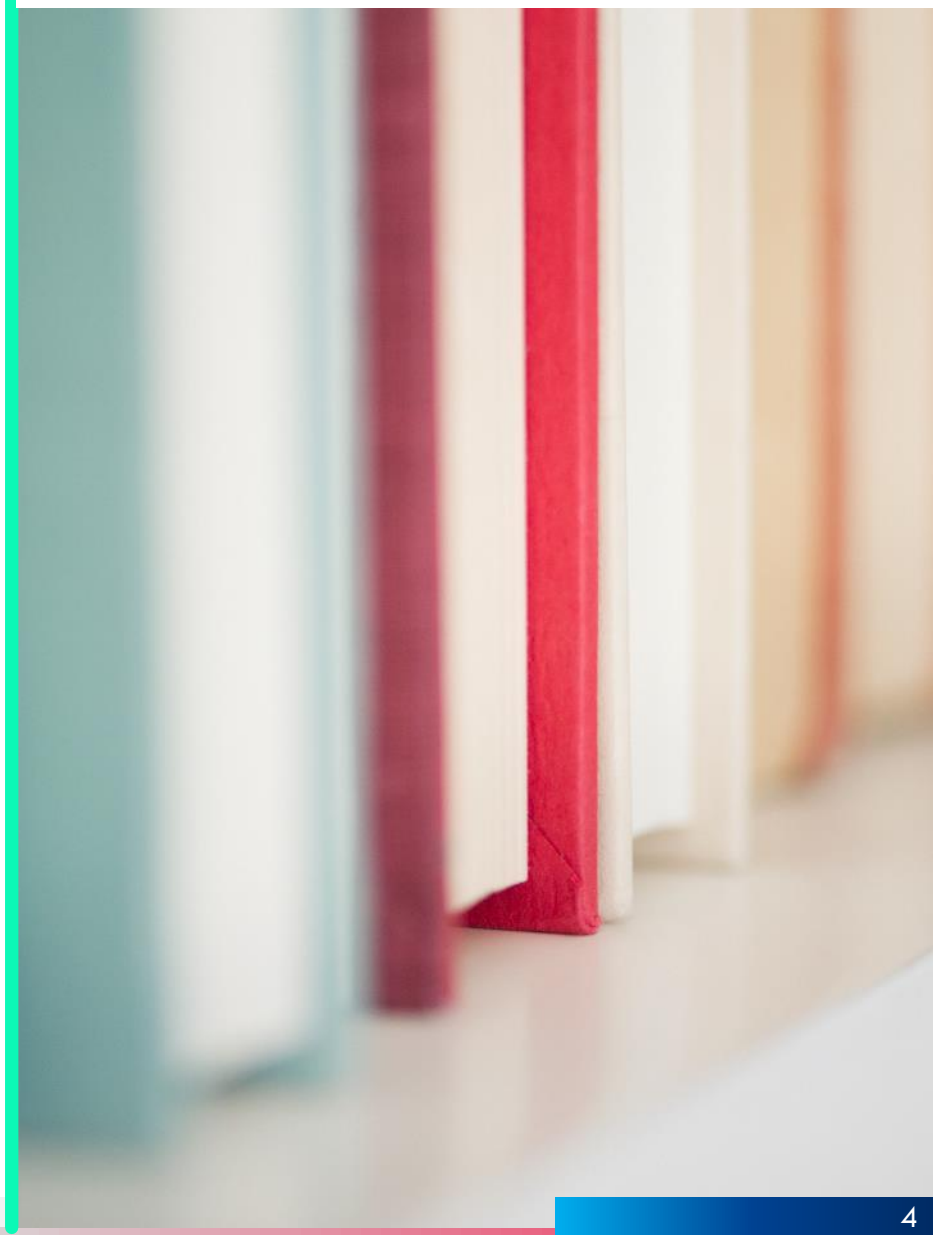
Quora is place and platform to:

- Ask questions
- Share knowledge about every thing
- Empower people learning from each other

100 million of people visiting Quora per month.

They ask similarly worded questions, this can cause:

- Seekers spends more time in searching the best answer
- Make writers feel they need to answer multiple versions of same question



Data set:

Data Consist of ‘404 290’ pair of questions:

	id	qid1	qid2	question1	question2	is_duplicate
1	0	1	2	what is the step by step guide to invest in share market in india	what is the step by step guide to invest in share market	0
2	1	3	4	what is the story of kohinoor diamond	what would happen if the indian government stole the kohinoor diamond back	0
3	2	5	6	how can i increase the speed of my internet connection while using a vpn	how can internet speed be increased by hacking through dns	0
4	3	7	8	why am i mentally very lonely how can i solve it	find the remainder when is divided by	0
5	4	9	10	which one dissolve in water quikly sugar salt methane and carbon di oxide	which fish would survive in salt water	0
6	5	11	12	astrology i am a capricorn sun cap moon and cap risingwhat does that say about me	i am a triple capricorn what does this say about me	1

The challenge consists of predicting the duplicate questions

Questions & Hypothesis:

1. How to drive a data science project given the data we have which conist of pair of questions
2. What kind of meaningful information we can extract from the two pair of question that can help in our prediction

Oscar to Reeva
15:01:37

I want to talk to you, I want to sort this out.. I don't want to have anything less than amazing for you and I.. I'm sorry for the things I say without thinking and for taking offense to some of your actions. The fact that I'm tired and sick isn't an excuse. I was upset that you just left me after we got food to go talk to a guy and I was standing tight behind you watching you touch his arm and ignore me and when I spoke up you introduced me which you could've done but when I left you just kept on chatting to him when clearly I was upset. I asked Martin to put on that kendrick lemar Album in the car and don't know it, granted that was a shut song but you could've just lent forward and whispered in my ear to change it scene I had to drive to pick up your friend. I was 30 min late and I know you don't like it when I drive fast but then you should've asked Gina to drive herself so that we wouldn't have to. When we left I was starving, the only good I'd had was a tiny wrap and everyone was leaving for lunch, I'm sorry I wanted to go but I was hungry and upset and although you knew it it wasn't like you came to chat to me when I left the table. I was upset when I left you cause I thought you were coming to me. I'm sorry I asked you to stop taping my neck yesterday, I know you were just trying to show me love.. I had a mad headache and should've just spoken to you softly. In sorry for asking you not to put on an accent last night.. Pretty much the same and didn't have the energy.

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

How to proceed?

Feature engineering and machine learning

- Transform the pair of questions to matrixes
- Calculate features
- Use machine learning to predict

Deep learning

- Transform the pair of questions to matrixes
- Use deep learning to predict

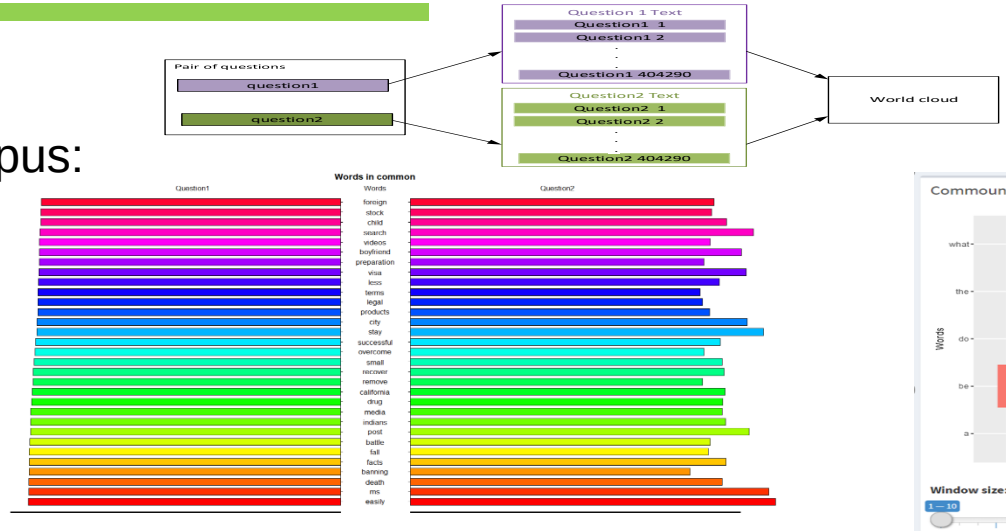
Agenda:

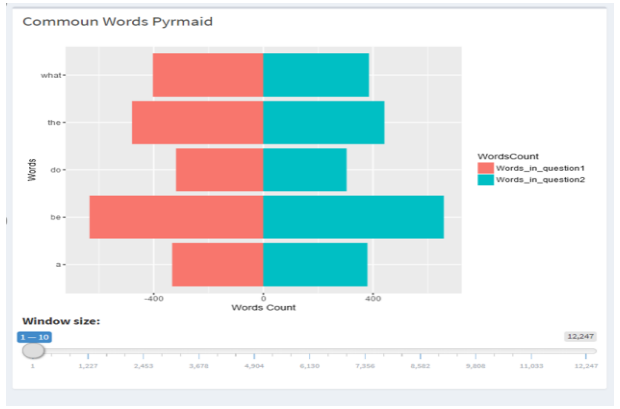
1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Data exploration on original data set:

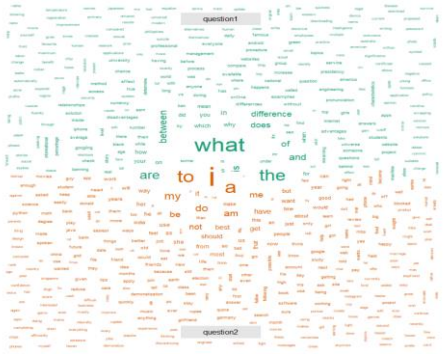
Data is pairs of two document corpus.

1. Comparison of words in two text corpus:
   - Word pyramid

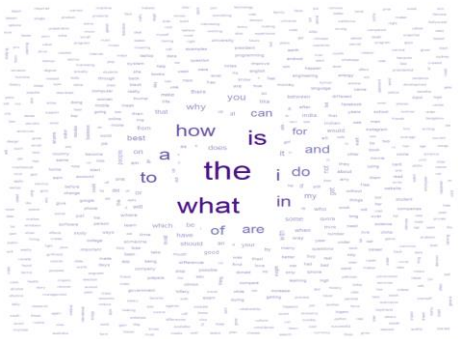




   - Word Cloud comparison



2. Common words Word Cloud:



Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Data Cleaning

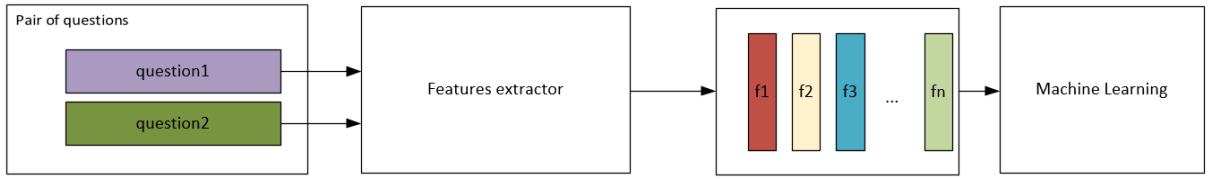
- tolower(): I like -> i like
- split_non_ascii(): 形から入る -> 形 か ら 入 る
- replace_appostrophs(): what's -> what is, isn’t -> is not
- remove_brackets(): "spiritual" -> spiritual
- remove_punctuations(): . ?

Keep as much as possible of words

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Feature engineering:



- Number of words
- Grammatical entities
- Sequence of grammatical entities and of words
- Word vector representation
- Cosine distance between two words

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Feature of words Statistics:

- Number of words in sentence
- Number of stop words sentence
- Number difference words between sentence (1 , 2), and (2,1)

Feature of Grammatical entities Statistics:

NLP ‘Part-of-Speech taggers’

- Number of verbs
- Number of adjectives
- Number of proper nouns

45 entities per one question.

We use PCA and Cosine Distance to reduce the dimensionality

Agenda:

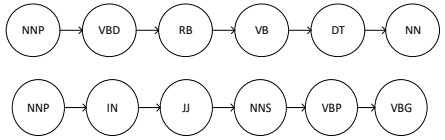
1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Features of Sequences:

TraminR is library in R used by social scientist to study careers trajectory
seqLLCS: length of the longest common subsequence of two sequences.
seqLLCP: the length of the longest common prefix of two sequences.

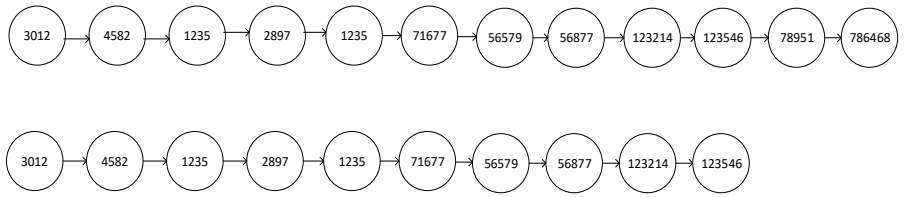
1. Sequence Grammatical entities:

- We get the sequence of grammatical entities using OpenNLP,
- We use the two functions of TraminR to quantify the sequence distances



2. Sequence of words in sentences:

- We extract vocabulary used in both questions, and we give an index for every word, Every question will be a sequence of index numbers.
- We calculate the distance of sequences of the pair of questions by using TraminR



Index	Word
3012	what
4582	is
1235	step
2897	by
71677	quide
56579	to
56877	invest
78951	in
123214	share
123546	market
786468	india

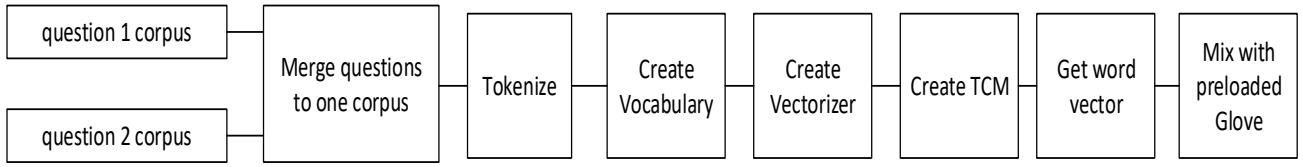
Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

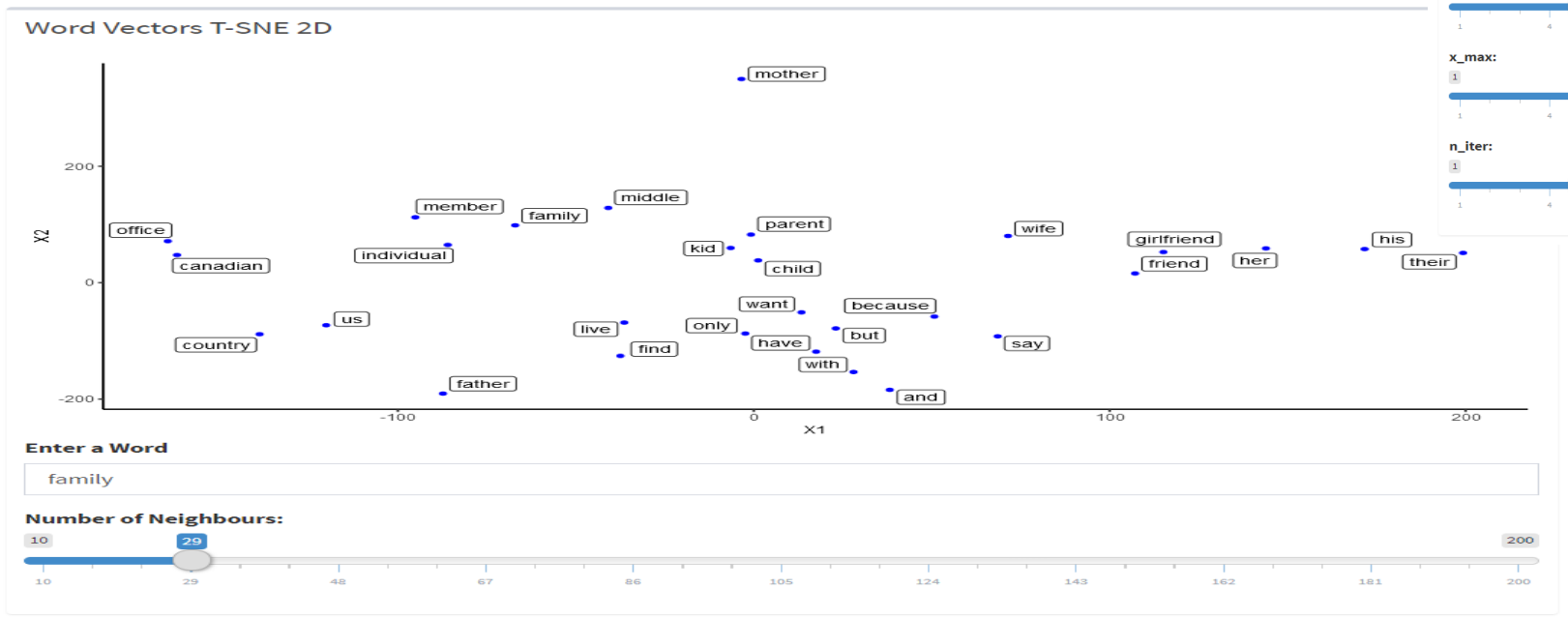
Features based on Word Vectors

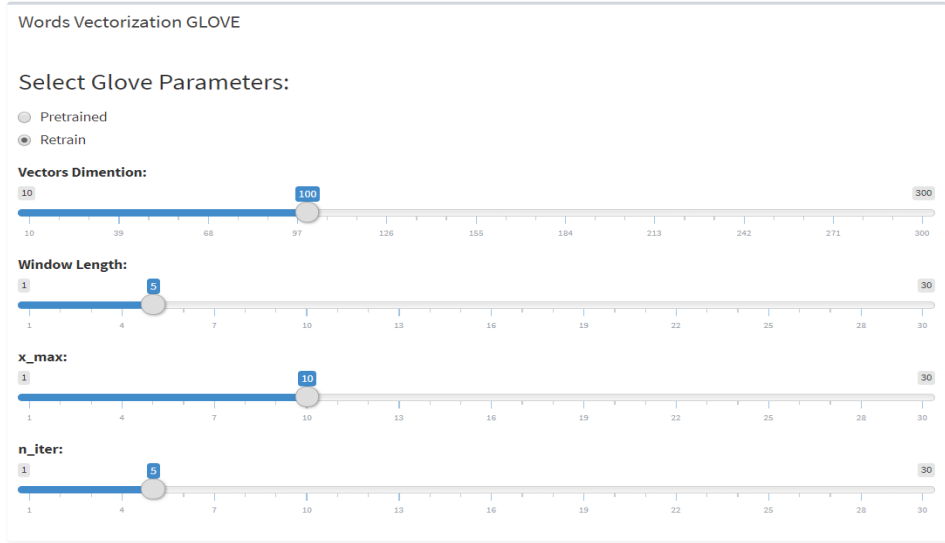
Word Vector:

We used Glove (Global vector) Algorithm to generate high dimension vectors for the words



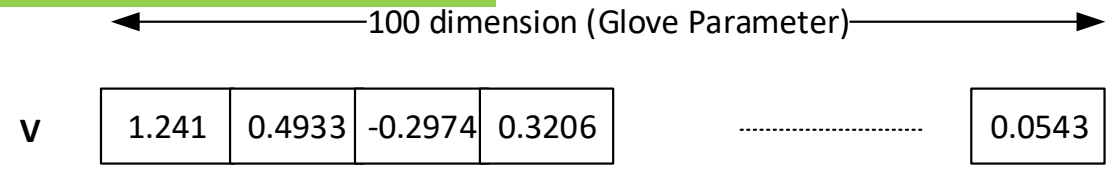
TCM: Term co-occurrence matrix



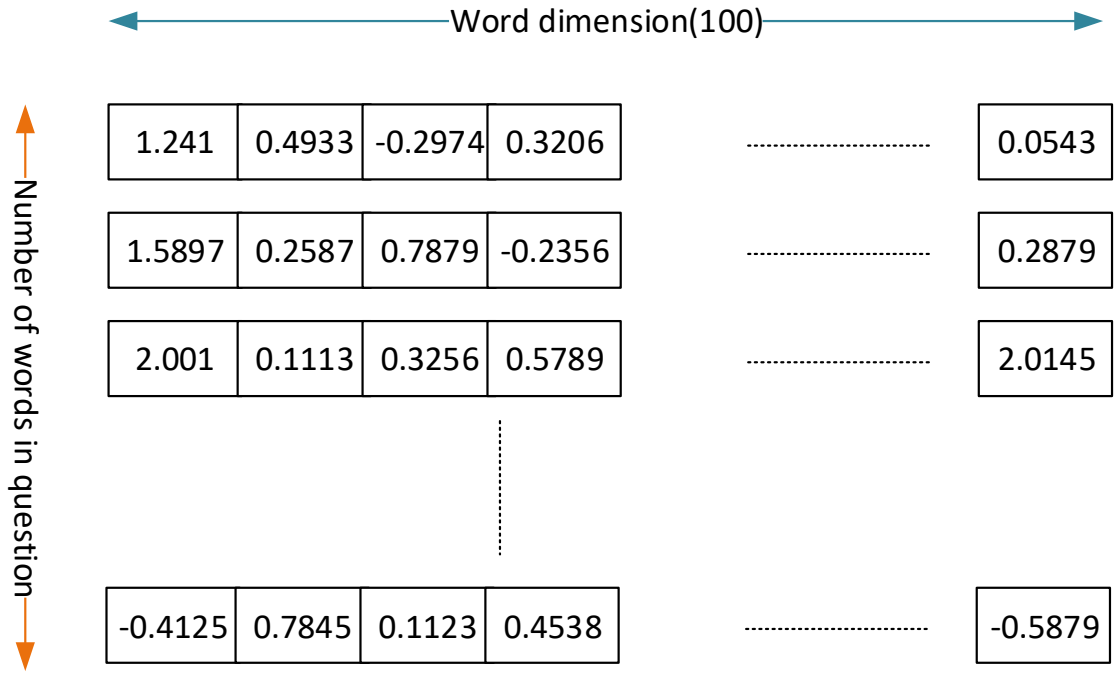


Matrix of sentence:

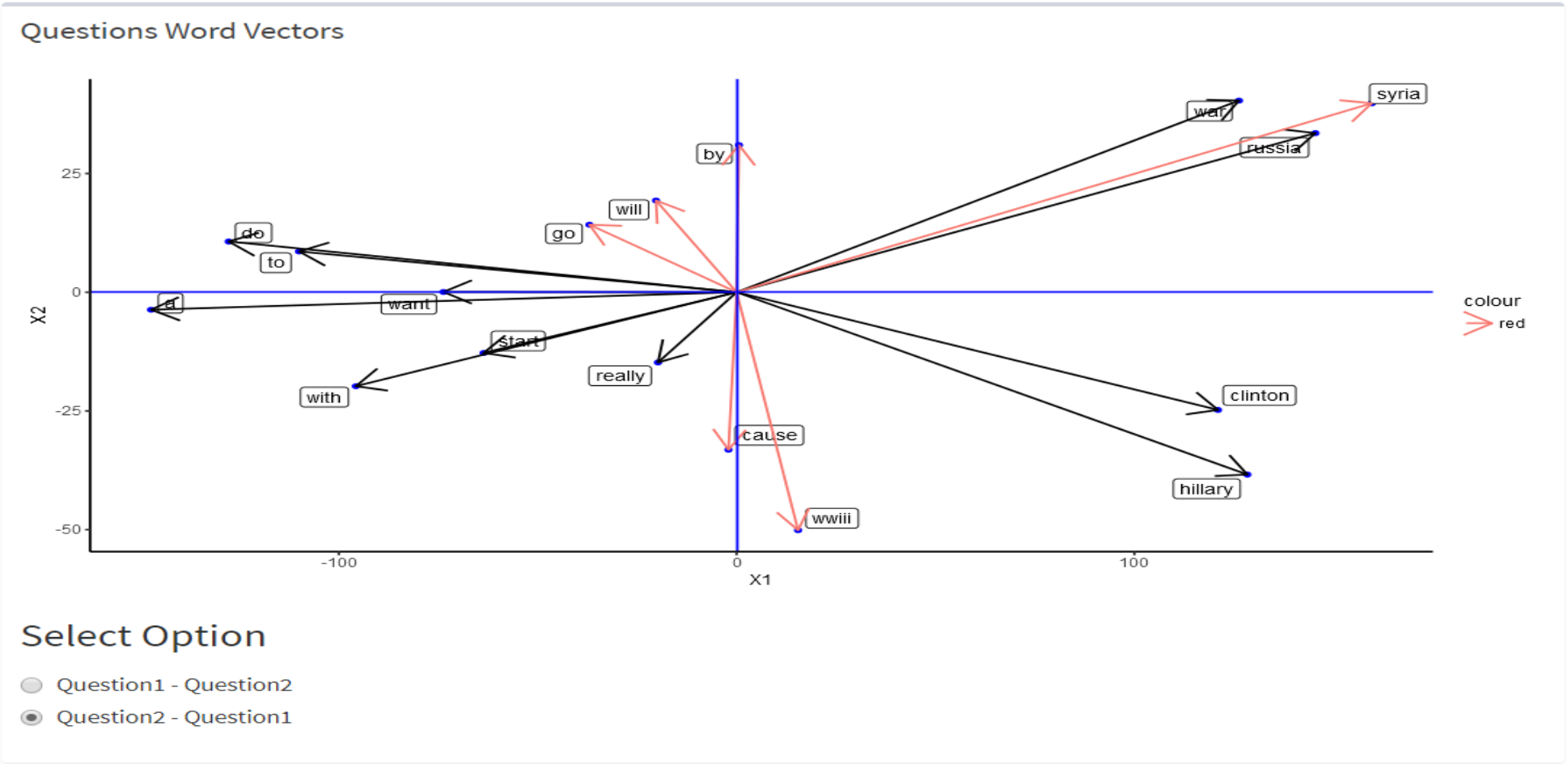
Word is Vector



Question sentence is Matrix

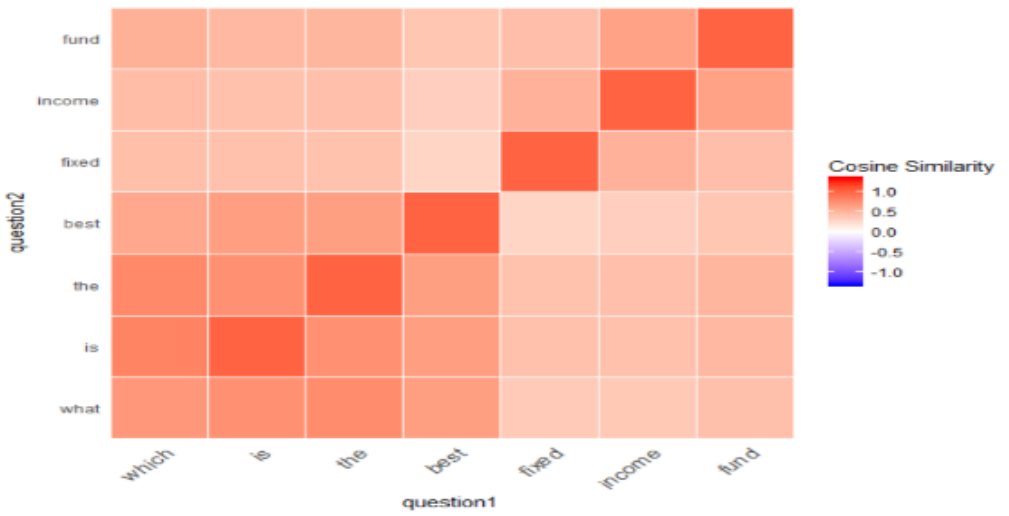


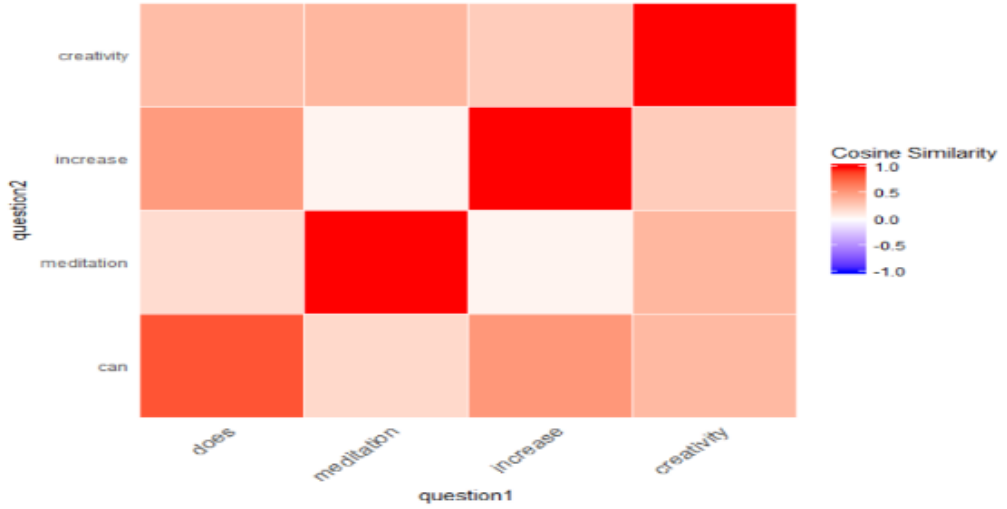
Cosine Distance between words of questions:

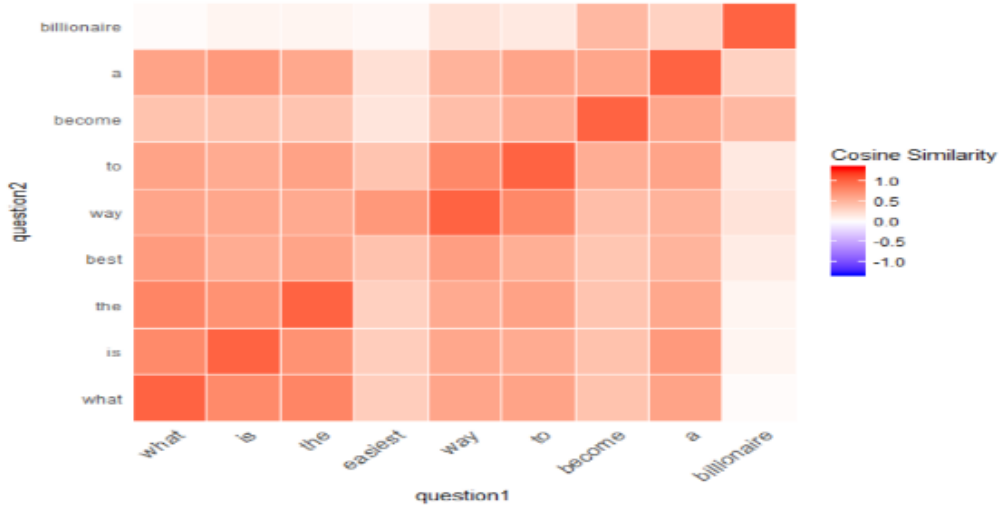


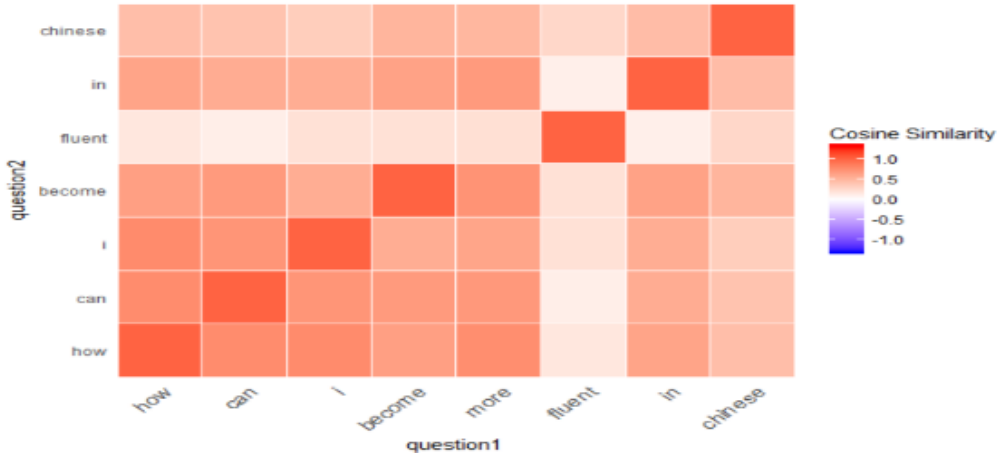
Similarity Matrix:

$$Cosine\ similarity(question1, question2) = \cos(\theta) = \frac{q1 * q2}{\|q1\|\|q2\|}$$

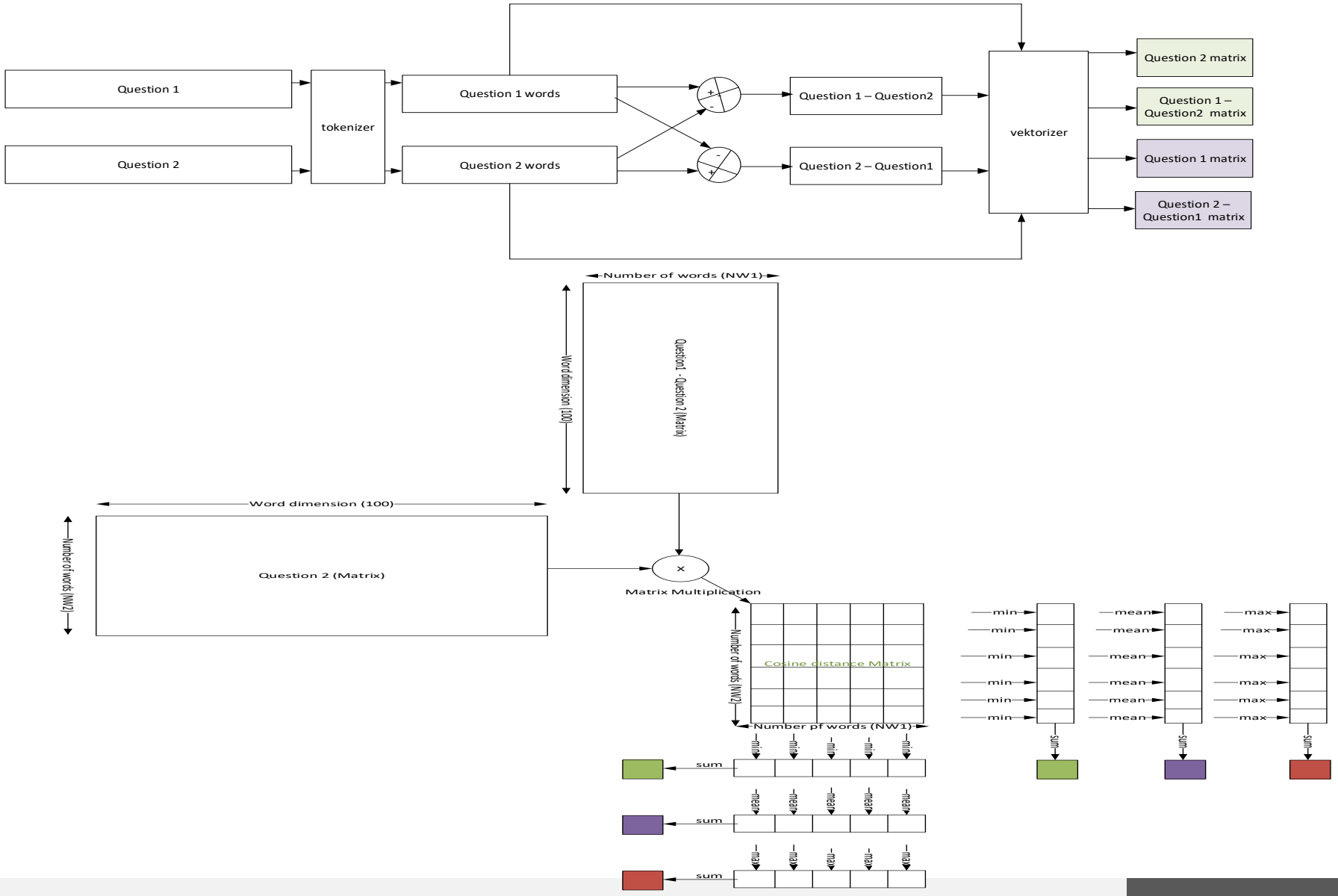








Calculate features from similarity matrixes:

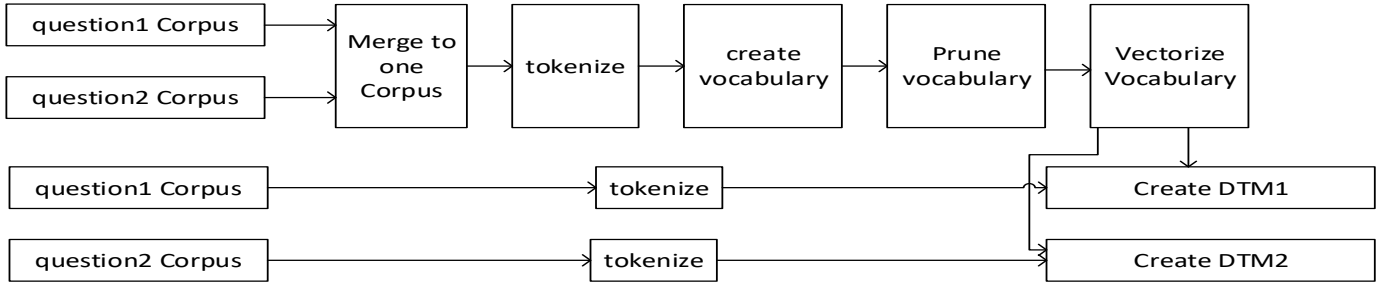


Agenda:

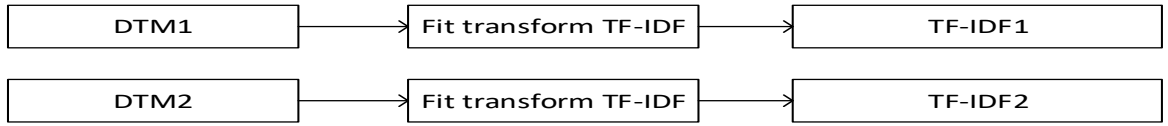
1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

Similarity distances using: DTM, TFIDF and LSA:

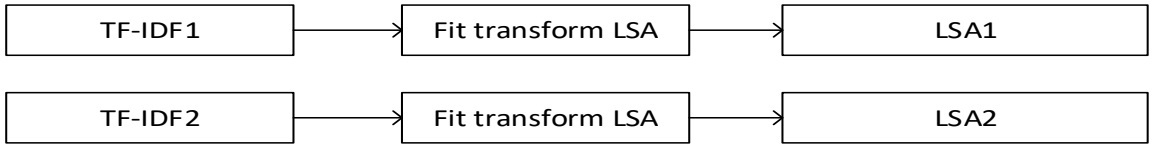
Matrixes pipelines:



DTM: Document term Matrix



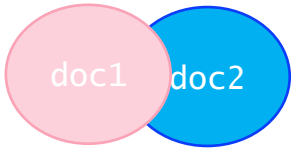
TF-IDF: Term Frequency Inverse Document Frequency



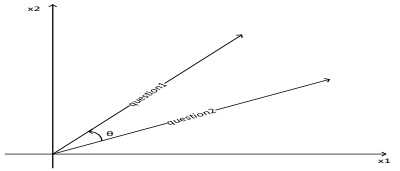
LSA: Latent Semantic Analysis

Distances

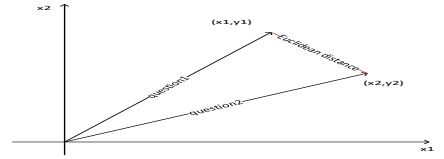
$$Jaccard\ distance(doc1, doc2) = \frac{doc1 \cap doc2}{doc1 \cup doc2}$$



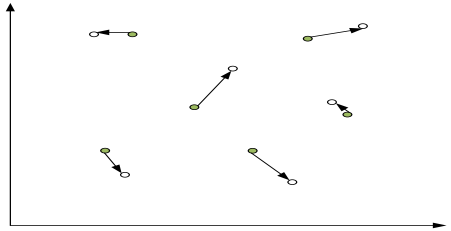
$$Cosine\ similarity(doc1, doc2) = \cos(\theta) = \frac{doc1\ doc2}{\|doc1\|\|doc2\|}$$



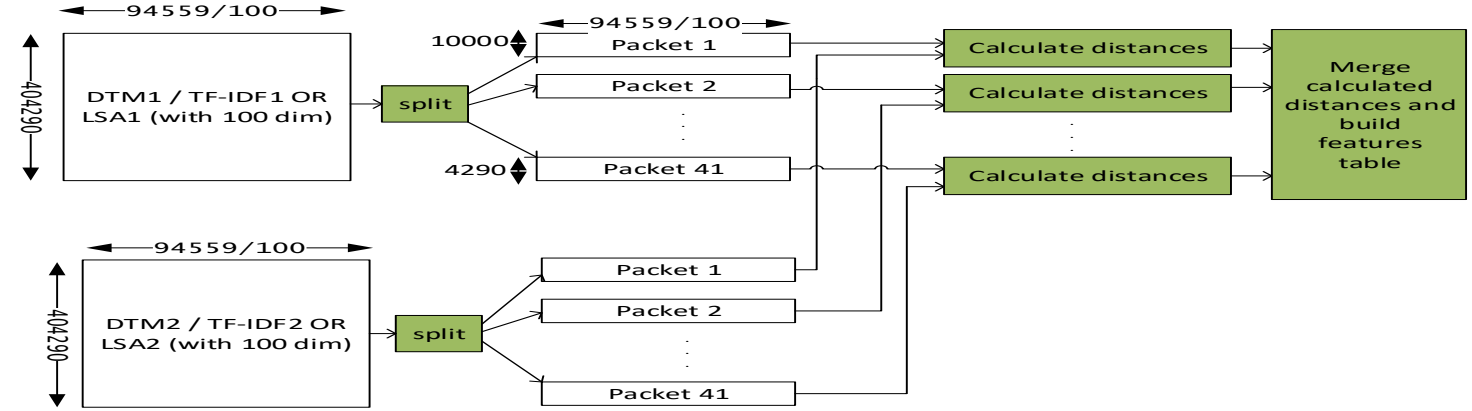
$$Euclidean\ distance = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

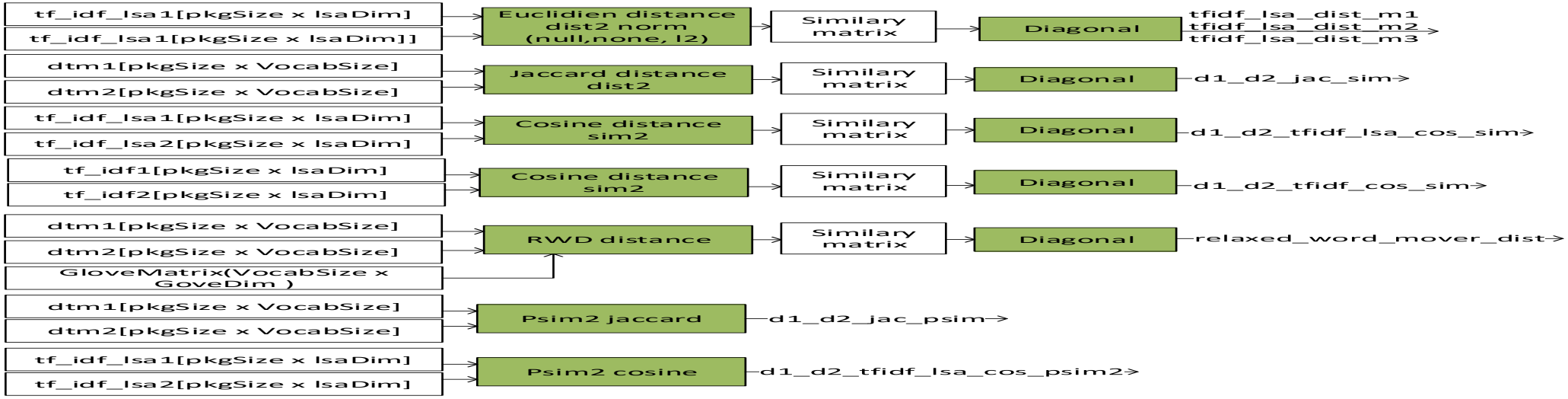


Relaxed Word Mover's Distance(WMD)



Features calculation pipelines:

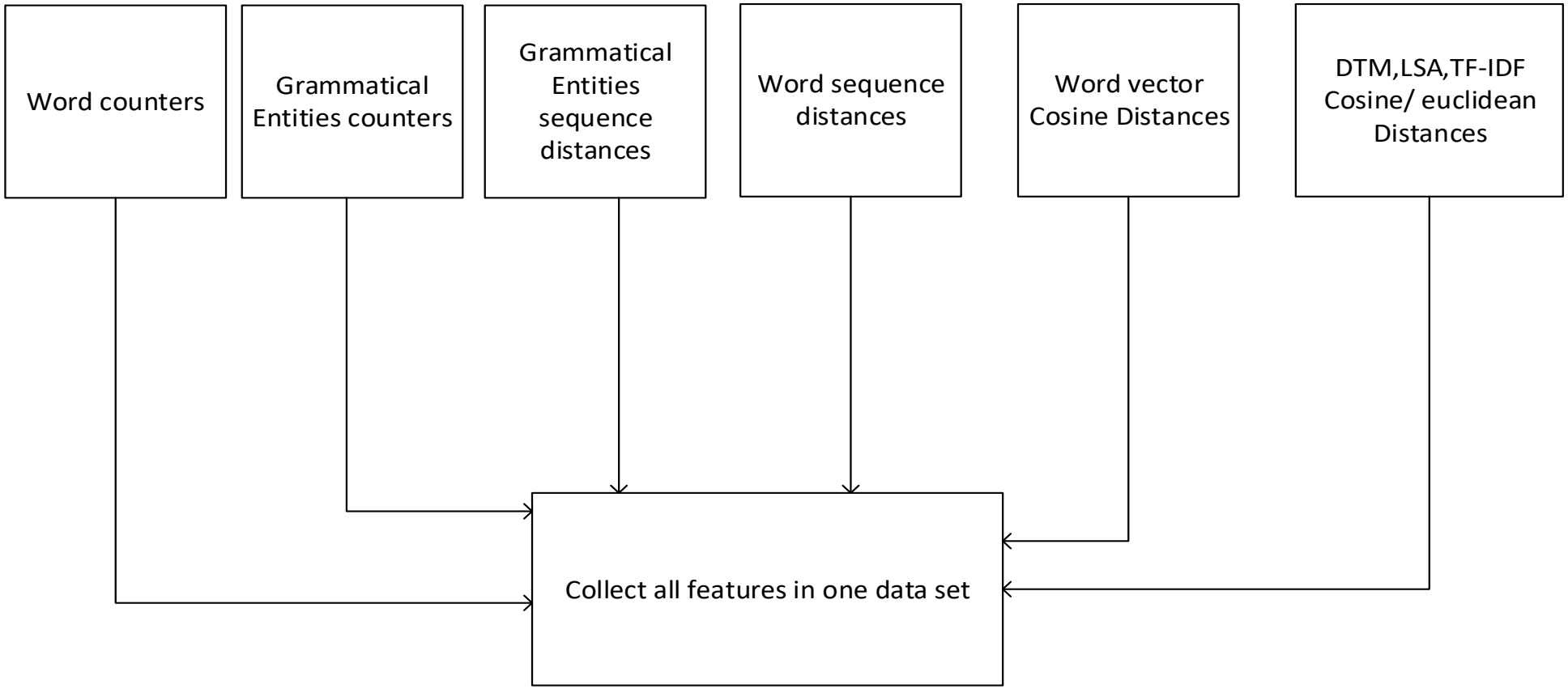




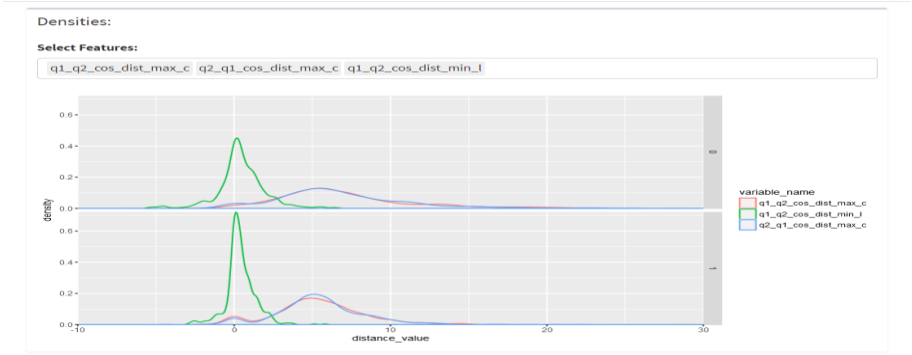
Agenda:

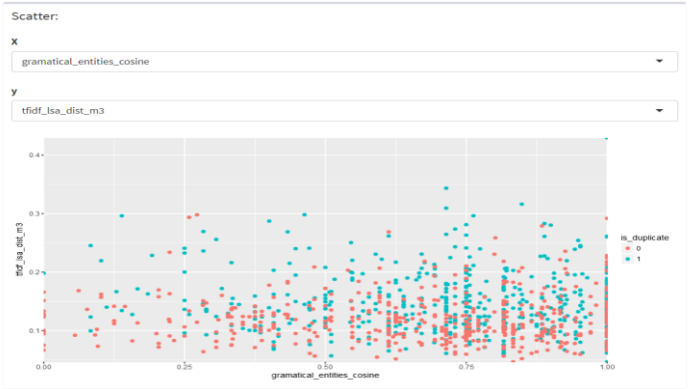
1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

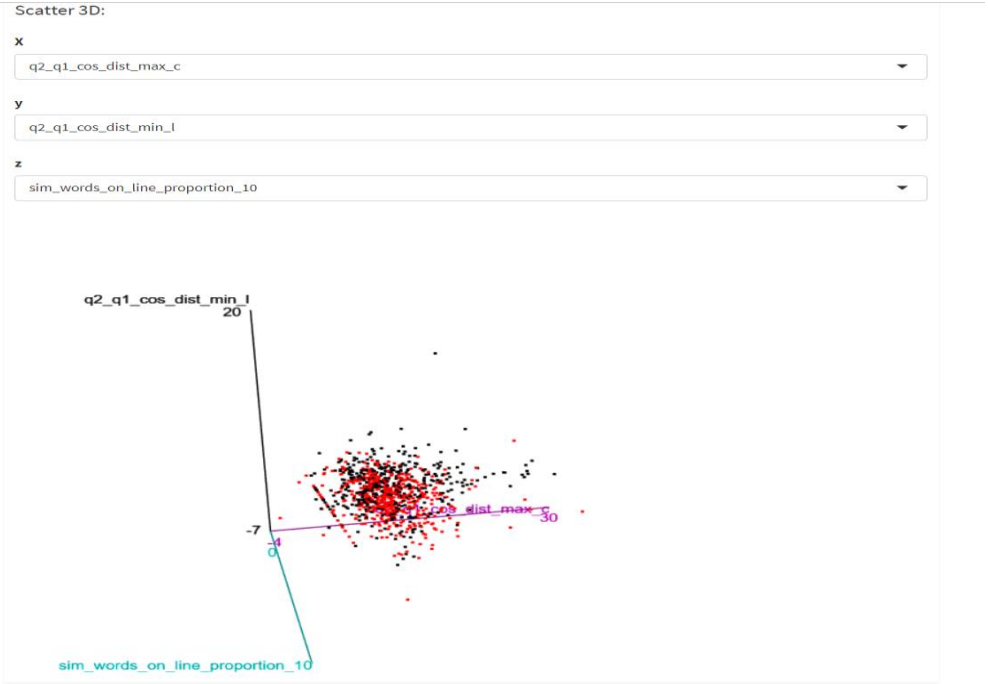
Merge calculated features



Visualization of features

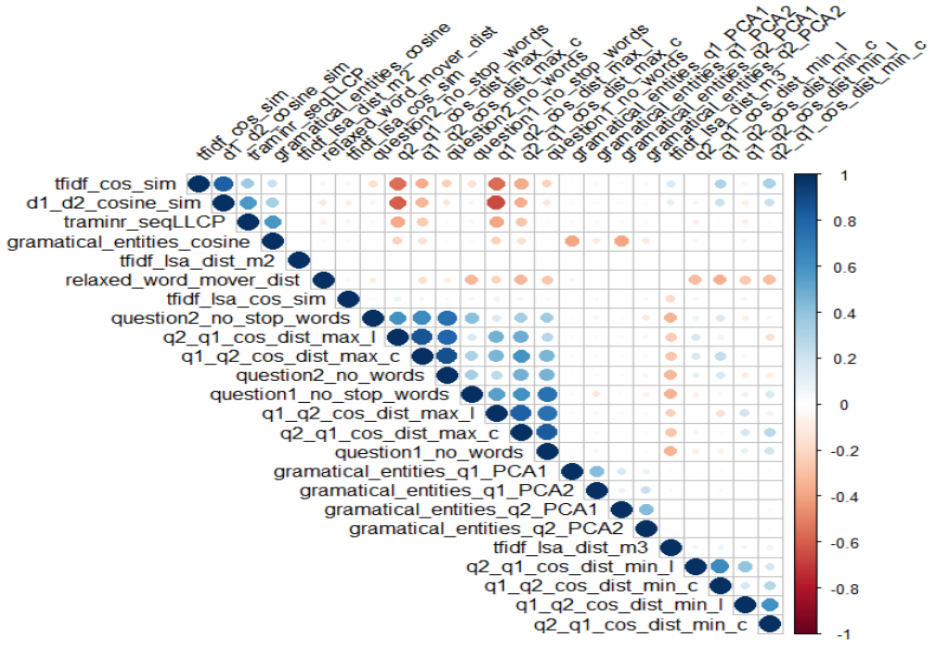






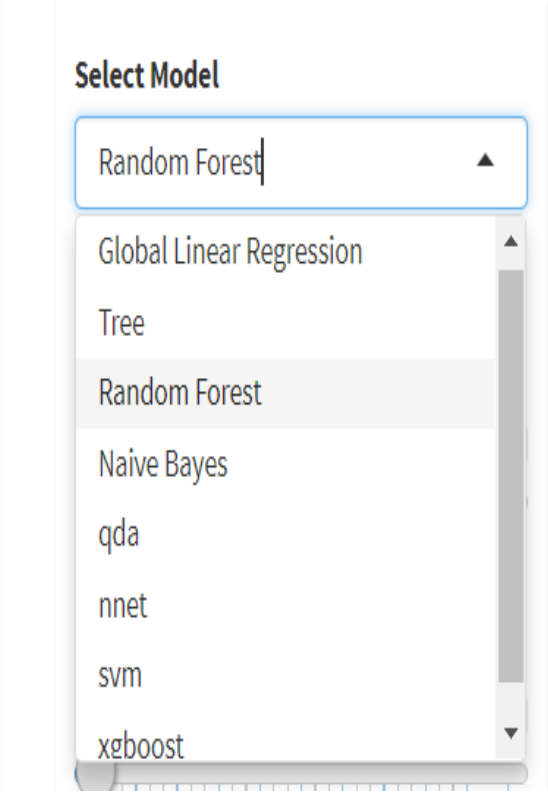
Remove redundancies:

We use corrplot to visualize the correlation between calculated features

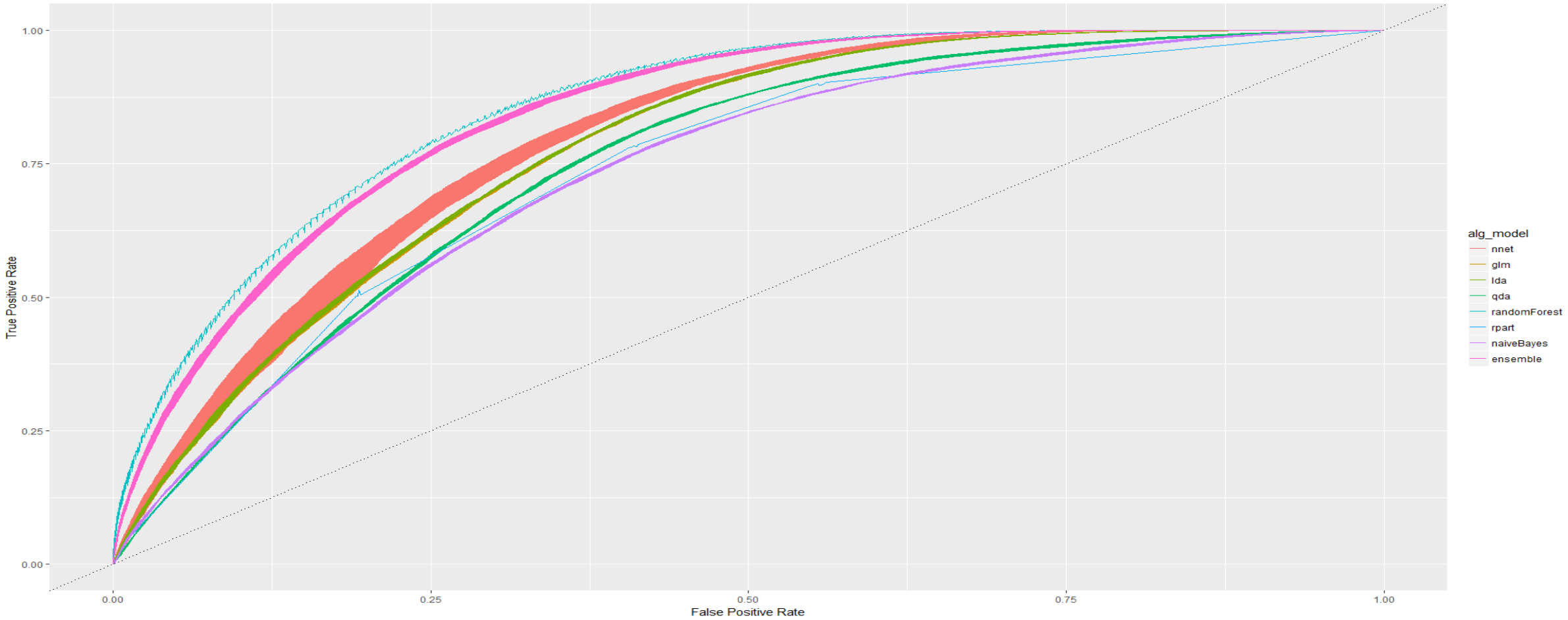


Machine Learning

Different Models:



Cross validation K-fold = 5 (ROC Curves)



Random-forest is the best with 78 % accuracy

Agenda:

1. Quora challenge
2. How to proceed?
3. Data exploration on original data set
4. Data cleaning
5. Features engineering and features extraction
   - Words statistics
   - Grammatical entities statistics
   - Grammatical entities and Words sequences
   - Cosine distances and similarity matrix
   - Similarity distances using: DTM, TFIDF and LSA
6. Apply Machine learning models to calculated Features
7. Deep Learning
   - Auto-encoders
   - Classifiers
8. Demo Shiny application

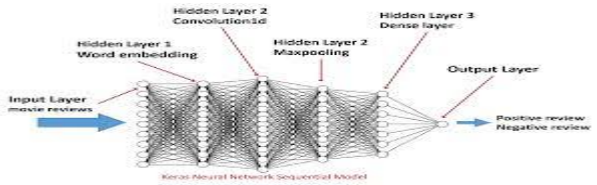
Deep Learning

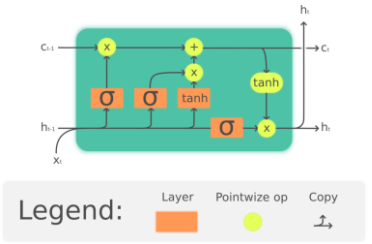
François Chollet

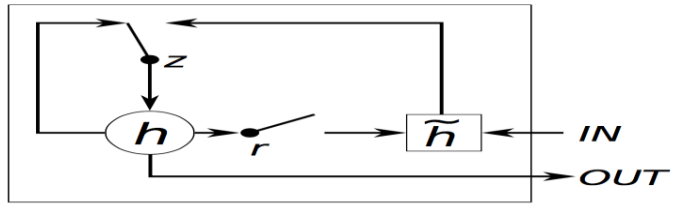
We used Keras library with backend TensorFlow

Deep Learning Layers:

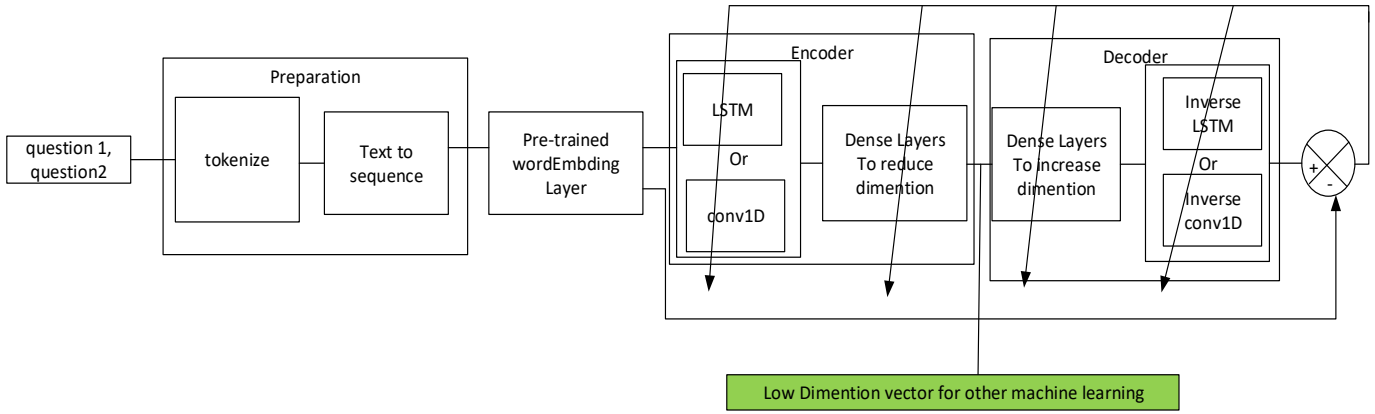
- LSTM, GRU: for word sequences
- DOT, Concatenate: Merging two inputs
- Dense Layers with activation functions: Relu, Sigmoid
- Drop-Out (Overfitting)
- Embedding
- CNN: Convolution Neural Network



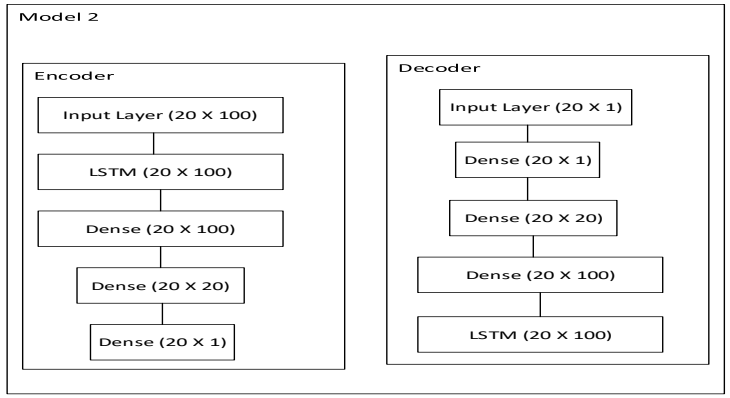


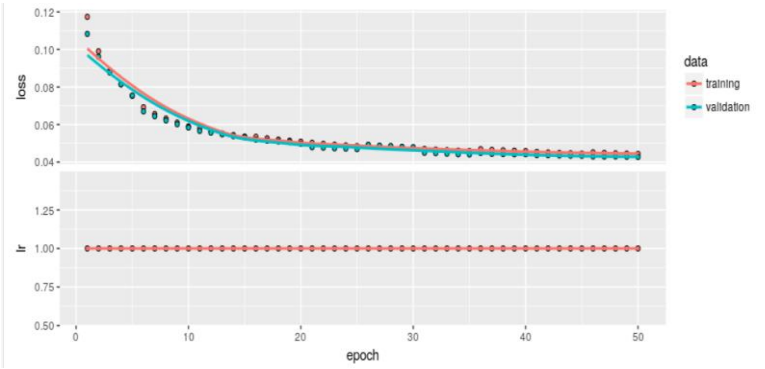


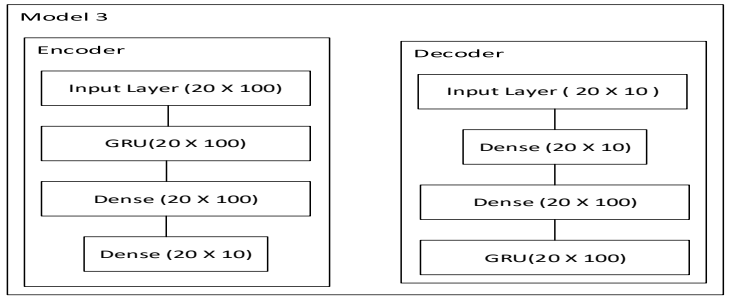
Auto-encoders:

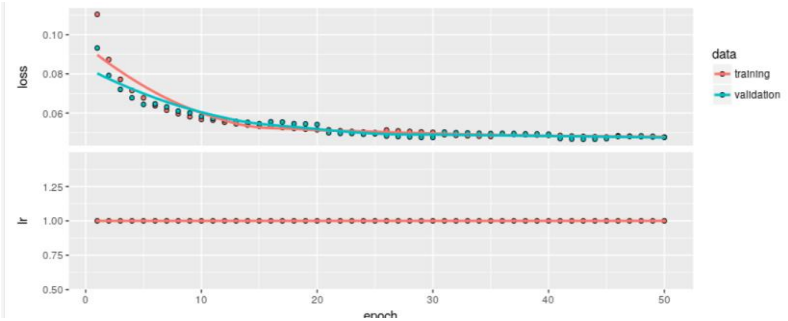


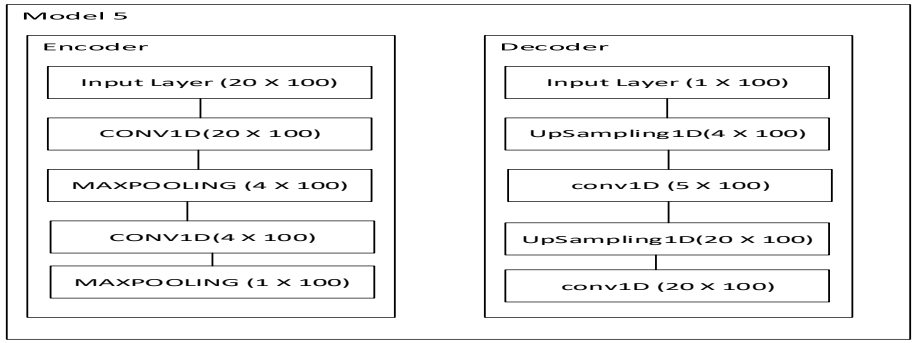
Auto-encoders different models:

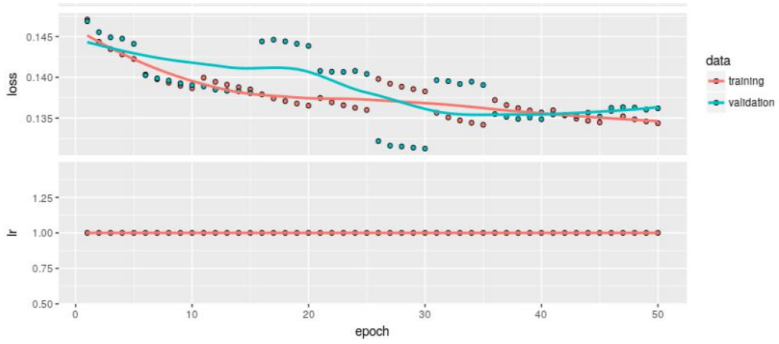




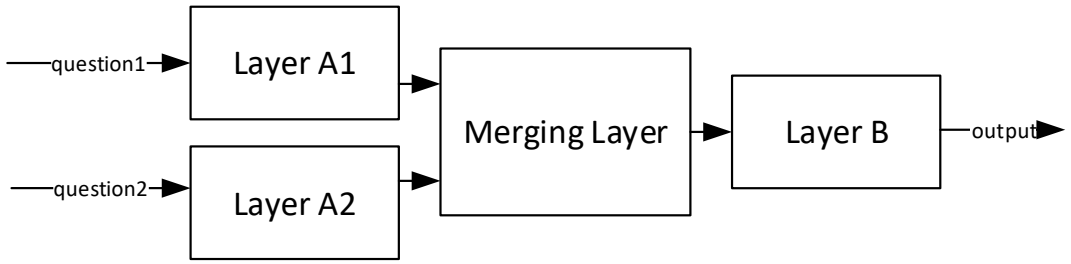




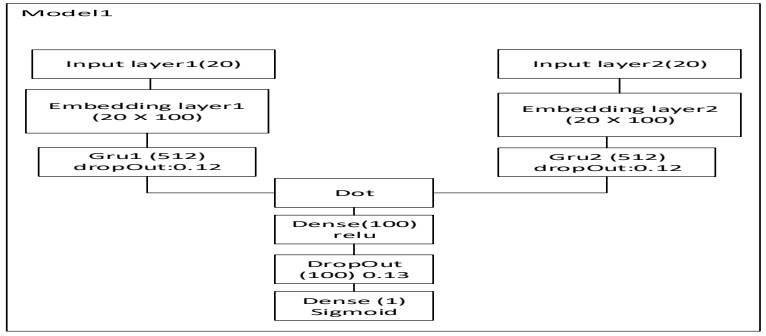


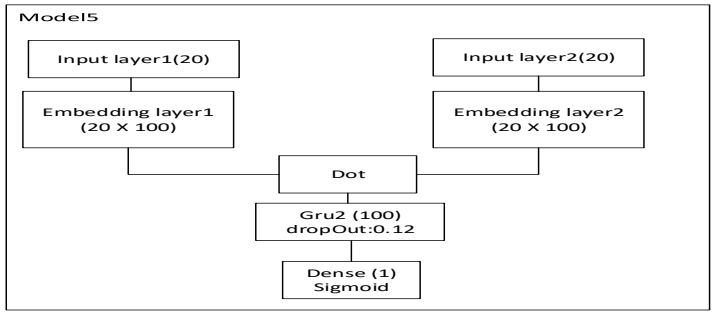


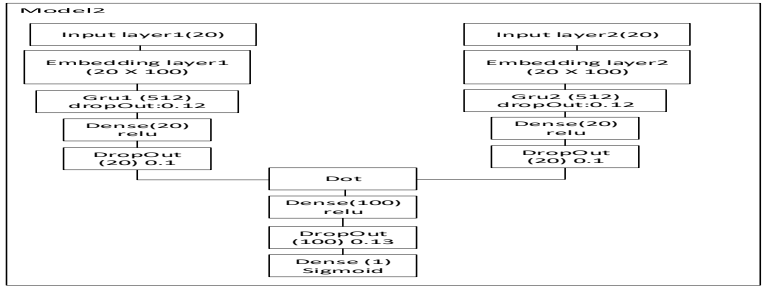
Deep Learning Classifiers

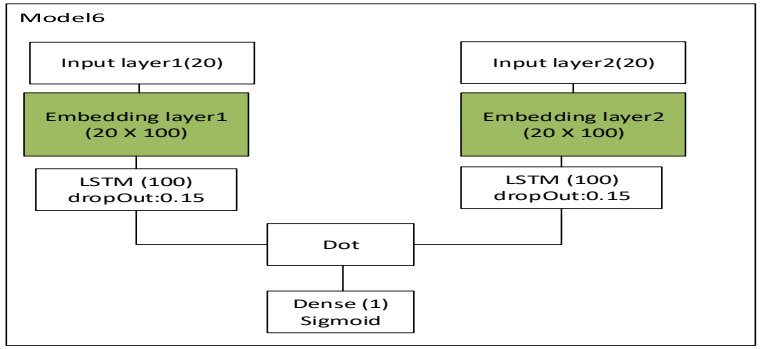


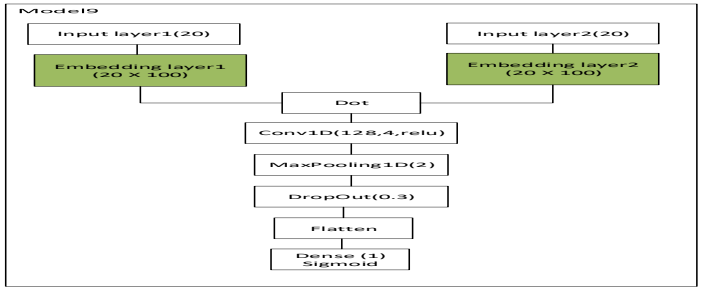
Deep Learning Classifier Models:

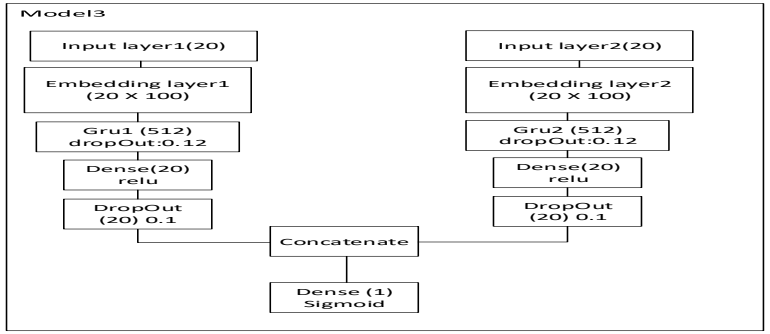


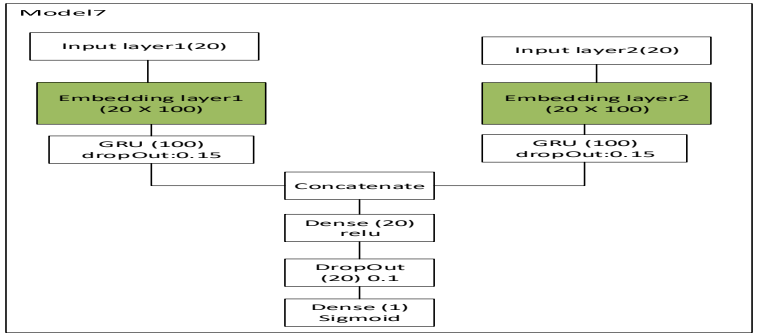


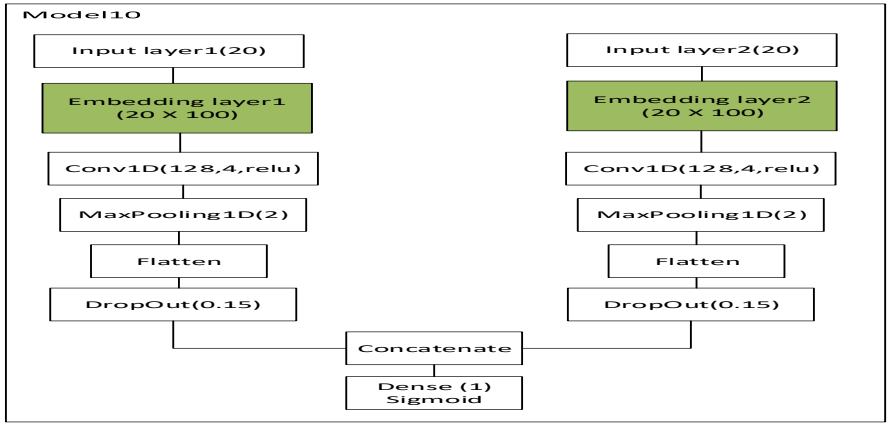


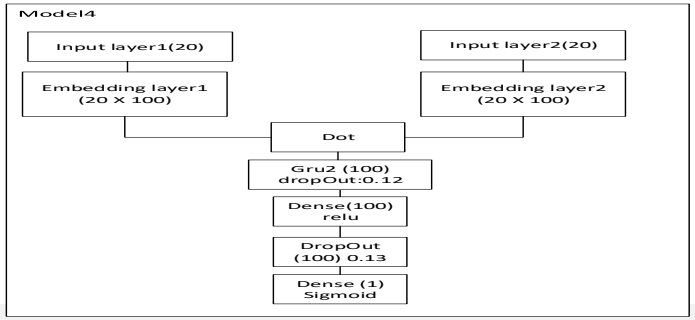


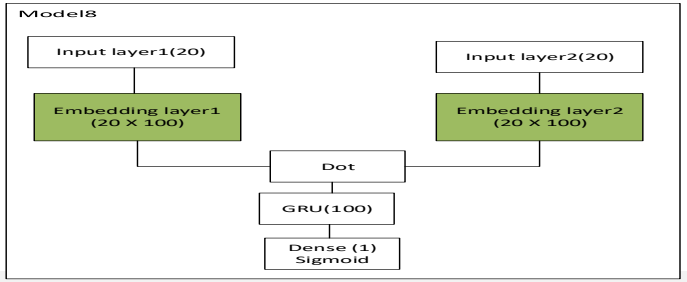




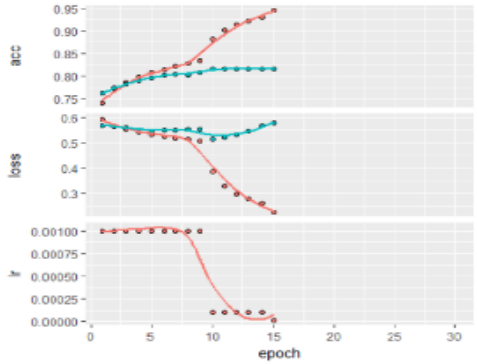


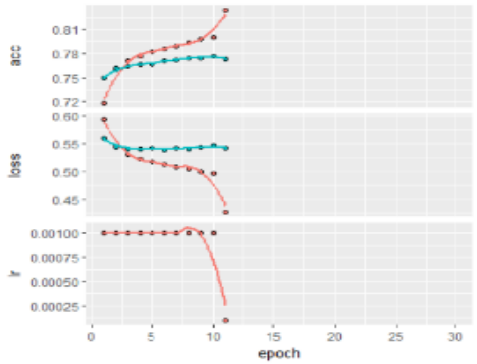


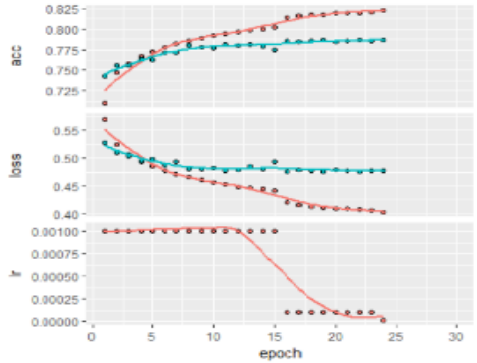


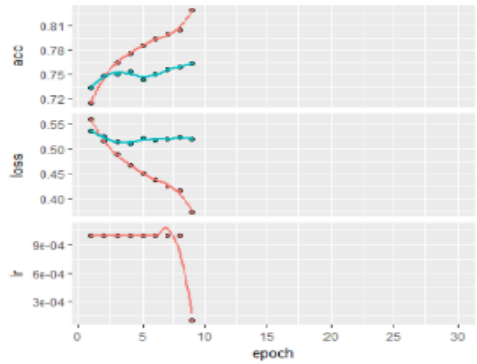


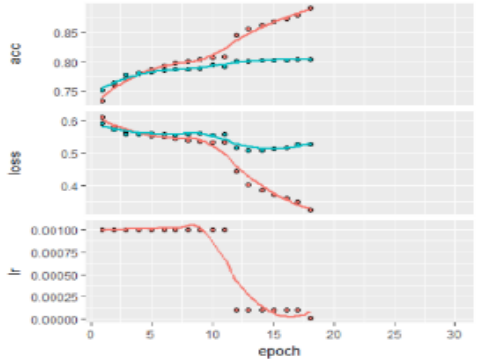
Accuracy and Loss curves of Classifier Models

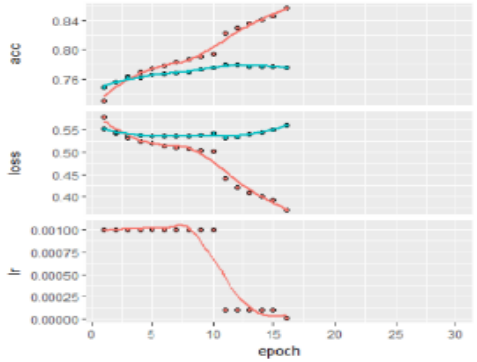


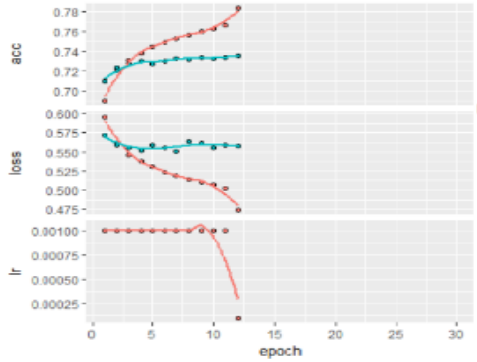


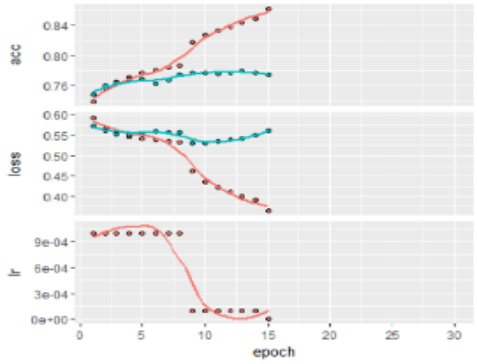


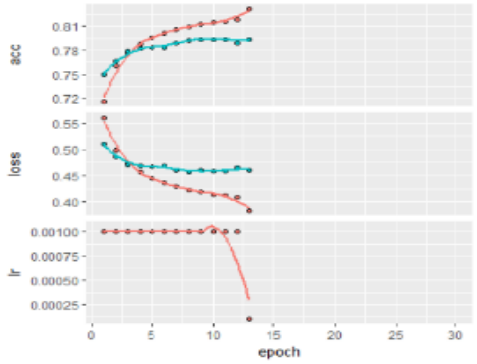


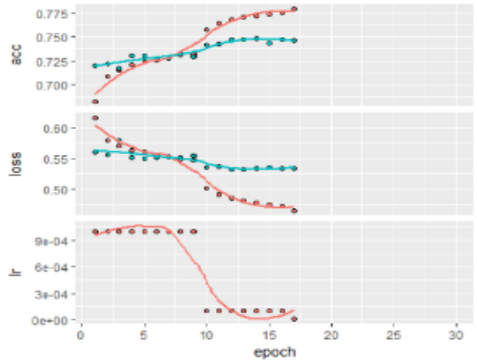




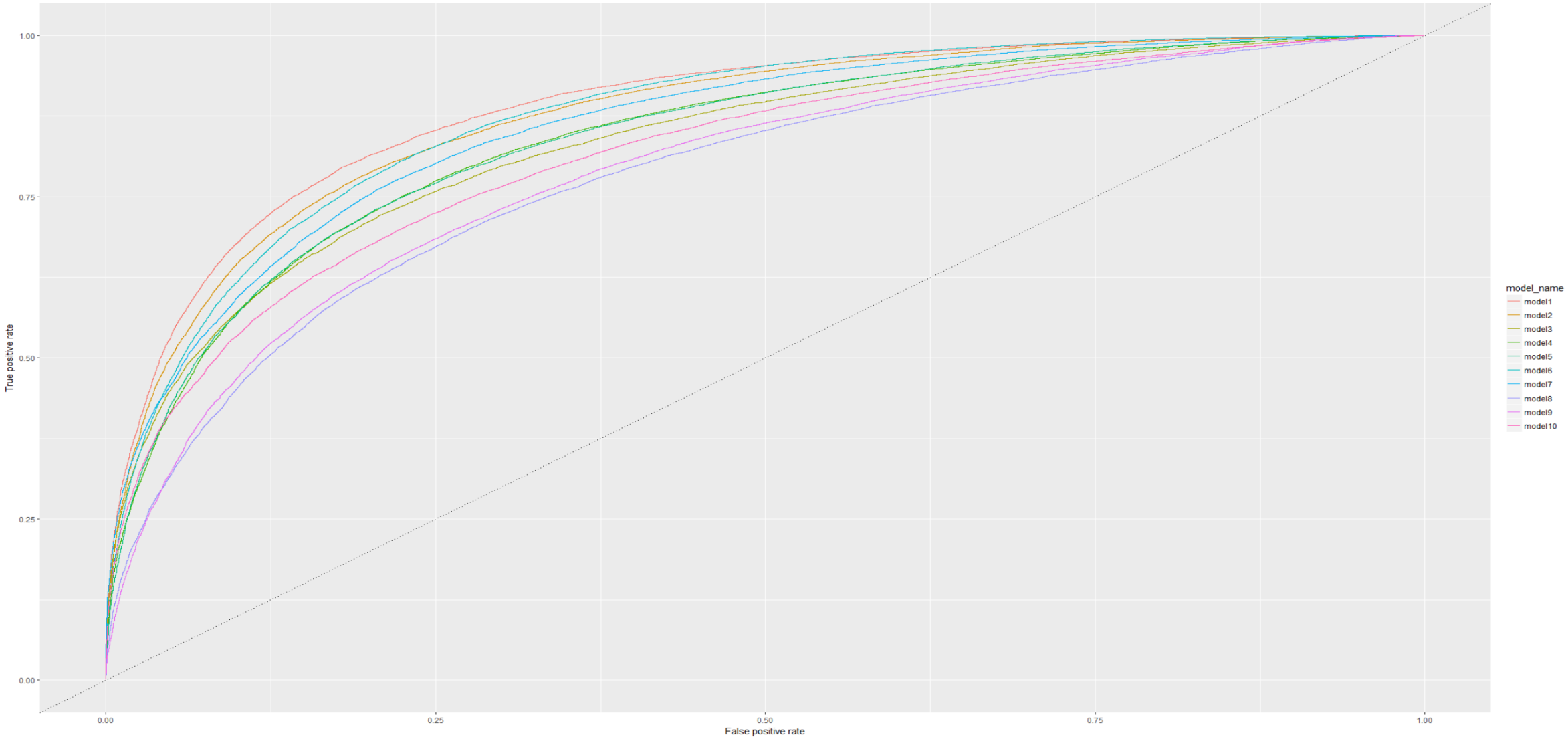








Compare the models with ROC Curves



Shiny Dashboard application:

Demo

Thank You

Abdelhak NEZZARI

abdelhak.nezzari@gmail.com