



# News Summarization using Feature Selection, Word2Vec and BERT Presentation:

Mohith Ankem  
CS256-Fall 2023

# Dataset:

- ❑ Which dataset (topic) are you working on?
- ❑ **BBC News Summary: Sport**
- ❑ Link: ***BBC News Summary/News Article/sport***
- ❑ What are the top words of your assigned dataset?  
**the, to, in, and, of, he, on, is ,it**

# Loading the Data:

- ❓ Initially, data was randomly being loaded into corpus.
- ❓ Have corrected data load to happen sequentially.

✓ How many txt files are under your assigned folder?

```
[9] print('There are {} text files under my assigned folder'.format(len(corpus)))
```

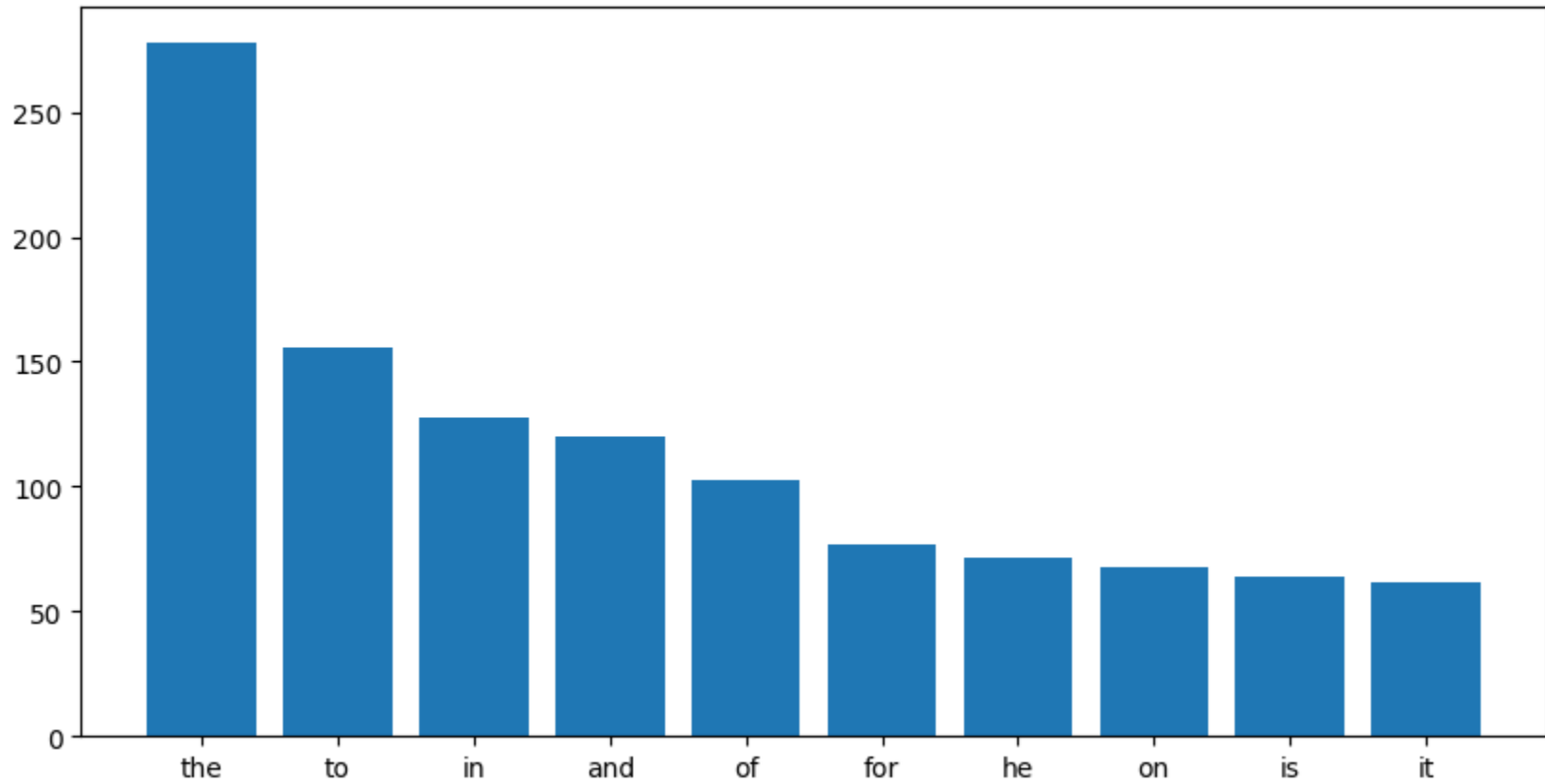
There are 511 text files under my assigned folder

10 rows x 10571 columns

```
[13] n_rows, n_columns = tfidf_df.shape  
print(f" Resultant vectorized dataframe has: {n_rows} rows x {n_columns} columns")
```

Resultant vectorized dataframe has: 2830 rows x 10571 columns

# Visualizing Top words:



Potential Stopwords:  
['the', 'to', 'in', 'and', 'of', 'for', 'he', 'on', 'is', 'it', 'was', 'with', 'at', 'but', 'his', 'we', 'that', 'has', 'have', 'will', 'be', 'as', 'said', 'not', 'england', 'after', 'from', 'by', 'out', 'they', 'against', 'win', 'are', 'beer

# Data preparation:

- Stopwords removal
- Lemmatization
- Stemming

# Part A: Milestone 1: Feature scores

## ❓ **Title features (F1):**

It is defined as a ratio of the number of matches of the Title words ( $Tw$ ) in the current sentence ( $S$ ) to the number of words ( $w$ ) of the Title ( $T$ )

## ❓ **Sentence Length (F2):**

It is defined as a ratio of the number of words ( $w$ ) in the current sentence ( $S$ ) to the number of words in the longest sentence ( $LS$ ) in the text.

## ❓ **Sentence position (F3):**

Determines how important a sentence is based on its position from top / bottom.

## ❓ **Term Weight (F5):**

It is defined as a ratio of the sum of the frequencies of term occurrences ( $TO$ ) in a sentence ( $S$ ) to the sum of the frequency of term occurrences in the text.

## ❓ **Proper Noun (F6):**

It is defined as a ratio of the number of proper nouns ( $PN$ ) in a sentence ( $S$ ) to the length ( $L$ ) of a sentence

## ❓ **Numerical Data (F7):**

It is defined as a ratio of the number of numerical data ( $ND$ ) in the sentence ( $S$ ) to the length ( $L$ ) of the sentence

# Part A: Milestone 1: Combined feature scores

Title of document'509': Melzer shocks Agassi in San Jose

Combined Feature Score for 'Second seed Andre Agassi suffered a comprehensive defeat by Jurgen Melzer in the quarter-finals of the SAP Open.' in document 509 is: 0.38753252982311814  
Combined Feature Score for 'Agassi was often bamboozled by the Austrian's drop shots in San Jose, losing 6-3 6-1.' in document 509 is: 0.44274689094333164  
Combined Feature Score for 'Defending champion and top seed Andy Roddick rallied to beat Sweden's Thomas Enqvist 3-6 7-6 (8-6) 7-5.' in document 509 is: 0.2882671423210433  
Combined Feature Score for 'But unseeded Cyril Saulnier beat the fourth seed Vincent Spadea 6-2 6-4 and Tommy Haas overcame eighth seed Max Mirnyi 6-7 (2-7) 7-6 (7-3) 6-2.' in document 509 is: 0.3352756991321123  
Combined Feature Score for 'Melzer has now beaten Agassi in two of their three meetings.' in document 509 is: 0.28505385026737967  
Combined Feature Score for '"I had a good game plan and I executed it perfectly," he said.' in document 509 is: 0.08666695417169801  
Combined Feature Score for '"It's always tough to come out to play Andre.' in document 509 is: 0.08139881177645883  
Combined Feature Score for '"I didn't want him to play his game.' in document 509 is: 0.06452540106951872  
Combined Feature Score for 'He makes you run like a dog all over the court.'" in document 509 is: 0.08436720142602497  
Combined Feature Score for 'And Agassi, who was more than matched for power by his opponent's two-handed backhand, said Melzer was an example of several players on the tour willing to take their chances against him.' in document 509 is: 0.215  
Combined Feature Score for '"A lot more guys are capable of it now," said the American.' in document 509 is: 0.09079968508615568  
Combined Feature Score for '"He played much better than me.' in document 509 is: 0.12247326203208558  
Combined Feature Score for 'That's what he did both times.' in document 509 is: 0.11981283422459894  
Combined Feature Score for '"I had opportunities to loosen myself up," Agassi added.' in document 509 is: 0.18617439200186003  
Combined Feature Score for '"But I didn't convert on the big points.'" in document 509 is: 0.21407754010695187

Title of document'510': Mirza makes Indian tennis history

Combined Feature Score for 'Teenager Sania Mirza completed a superb week at the Hyderabad Open by becoming the first Indian in history to win a WTA singles title.' in document 510 is: 0.4006132302818878  
Combined Feature Score for 'In front of a delirious home crowd, the 18-year-old battled past Alyona Bondarenko of the Ukraine 6-4 5-7 6-3.' in document 510 is: 0.2705507380060445  
Combined Feature Score for 'Mirza, ranked 134 in the world, sunk to her knees in celebration after serving out the match against Bondarenko.' in document 510 is: 0.1822429976072792  
Combined Feature Score for '"It is a big moment in my career and I would like to thank everyone who has been a part of my effort," she said.' in document 510 is: 0.09363981272198901  
Combined Feature Score for '"This win has made me believe more in myself and I can now hope to do better in the coming days.' in document 510 is: 0.07513809941154376  
Combined Feature Score for '"I wanted to win this tournament very badly since it was in my hometown.'" in document 510 is: 0.06976891615541922  
Combined Feature Score for 'At the Australian Open in January, Mirza became the first Indian woman to reach the third round of a Grand Slam before losing to eventual champion Serena Williams.' in document 510 is: 0.20116093167701862  
Combined Feature Score for 'And a year ago, she became the youngest Indian to win a professional title by claiming the doubles at the Hyderabad Open.' in document 510 is: 0.12880049164569846  
Combined Feature Score for 'Mirza, playing in her first WTA final, began nervously in front of a raucous home crowd - committing three double faults in her opening service game.' in document 510 is: 0.1571666462167689  
Combined Feature Score for 'But from 0-2 down, Mirza broke serve twice in a row and held on to her advantage to take the first set.' in document 510 is: 0.17403690017596427  
Combined Feature Score for 'In a see-saw second set, Bondarenko raced into a 5-2 lead and though Mirza hauled herself level, the Ukrainian broke again before finally levelling the match.' in document 510 is: 0.1908893217007941  
Combined Feature Score for 'Mirza rediscovered the aggressive strokes that took her to the first set in the decider established a 5-2 lead.' in document 510 is: 0.2006754972354768  
Combined Feature Score for 'At 5-3, the stadium erupted in celebration when Mirza thought she had delivered an ace to secure victory but the serve was ruled to have clipped the net.' in document 510 is: 0.22796713825774884  
Combined Feature Score for 'Mirza eventually lost the point but to the relief of the crowd, she broke Bondarenko again in the next game to clinch the title.' in document 510 is: 0.27681399649430327

Title of document'511': Roddick to face Saulnier in final

Combined Feature Score for 'Andy Roddick will play Cyril Saulnier in the final of the SAP Open in San Jose on Sunday.' in document 511 is: 0.49203559139784947  
Combined Feature Score for 'The American top seed and defending champion overcame Germany's Tommy Haas, the third seed, 7-6 (7-3) 6-3.' in document 511 is: 0.28602292190808832  
Combined Feature Score for '"I was feeling horrible earlier in the week," Roddick said.' in document 511 is: 0.19883086021505375  
Combined Feature Score for '"I thought tonight was a step in the right direction.'" in document 511 is: 0.12778673835125448  
Combined Feature Score for 'Saulnier battled to a 6-7 (3-7) 6-3 6-3 win over seventh seed Jurgen Melzer, who twisted his ankle early in the second set.' in document 511 is: 0.3982203823178017  
Combined Feature Score for 'Roddick won the last four points of the first-set tie-break before being broken at the start of the second set.' in document 511 is: 0.15939493971977844  
Combined Feature Score for 'But he broke straight back and then broke Haas again to lead 4-2.' in document 511 is: 0.21160913978494622  
Combined Feature Score for '"It's extremely frustrating when you have chances against a top-five player and don't do anything with them," admitted Haas.' in document 511 is: 0.11195343434343434  
Combined Feature Score for '"I rushed a few backhands and he took advantage.'" in document 511 is: 0.06261114123068583  
Combined Feature Score for 'Saulnier will move into the world's top 50 for the first time after his passage through to the final.' in document 511 is: 0.26260395233943623  
Combined Feature Score for '"It's taken a lot of work and a lot of fighting in my mind," he revealed.' in document 511 is: 0.13816922683051716  
Combined Feature Score for '"Sometimes I didn't believe I could get to a final and now I am here.' in document 511 is: 0.19411418330773167  
Combined Feature Score for 'I've stayed mentally strong.' in document 511 is: 0.0991551459293395  
Combined Feature Score for '"I'm on the way.' in document 511 is: 0.10672811059907836  
Combined Feature Score for 'I'll keep fighting and work a lot and I'll be up there.'" in document 511 is: 0.21744669218151996

# Part A: Milestone 1:

- ❓ What is your proposed formula that combines the F scores for any given sentence?

```
f_total_score = (f1_score * 0.30) + (f2_score * 0.03) + (f3_score * 0.17) + (f5_score * 0.10) + (f6_score * 0.17) + (f7_score * 0.23)
return f_total_score
```



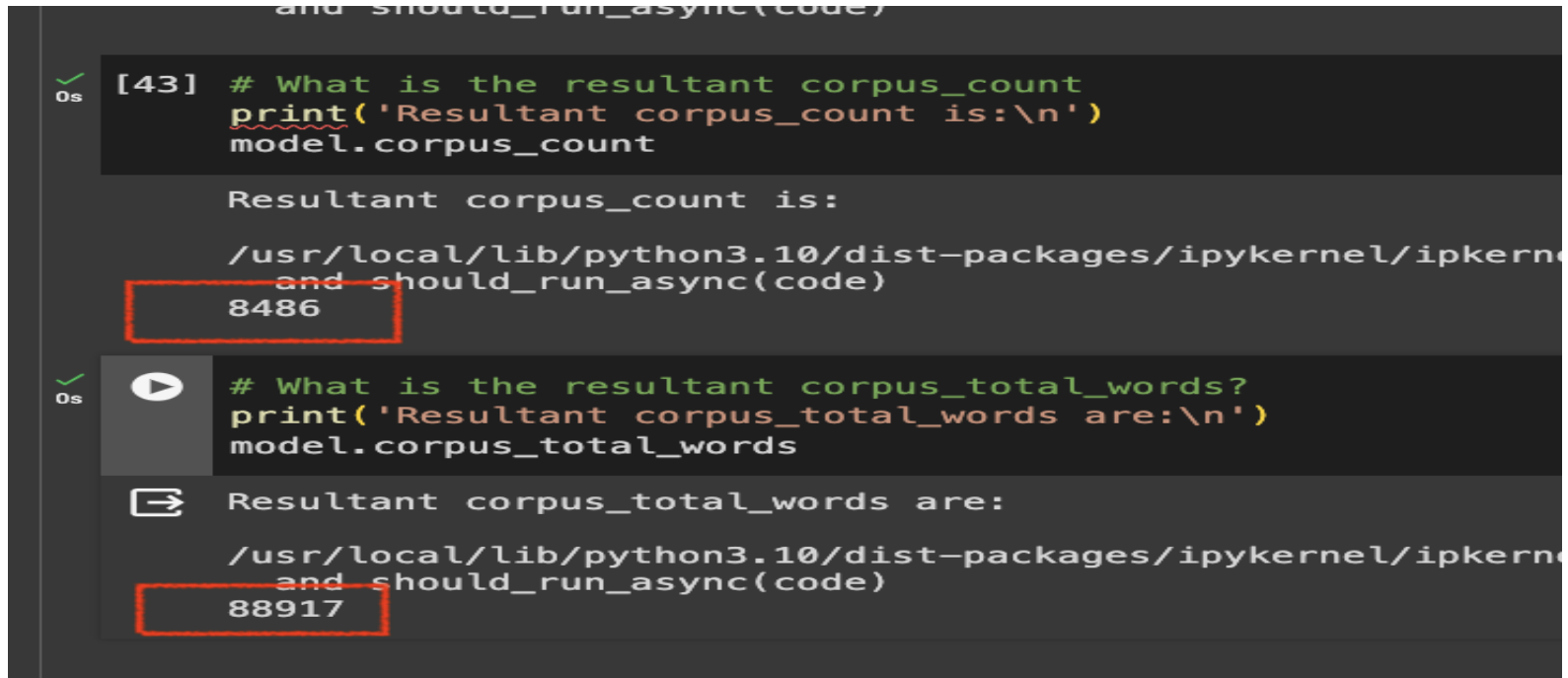
# Milestone 2: Data preparation / cleaning:

- Tokenization, Alphanumeric filtering, Stopword removal
- Make sure each sentence is in an array by itself:

```
['two', 'moments', 'magic', 'brian', 'guided', 'ireland', 'workmanlike', 'victory', 'italy']  
['pair', 'classic', 'outside', 'breaks', 'ireland', 'captain', 'set', 'tries', 'geordan', 'murphy', 'peter', 'stringer']  
['italy', 'led', 'early', 'second', 'half', 'stringer', 'try', 'gave', 'ireland', 'lead', 'never', 'lost']  
['hosts', 'cut', 'gap', '10', 'minutes', 'left', 'nearly', 'scored', 'ludovico', 'nitoglia', 'denis', 'hickie', 'try', 'ensured', 'irish', 'victory']  
['italy', 'came', 'flying', 'blocks', 'took', 'lead', 'luciano', 'orquera', 'penalty', 'seven', 'minutes']  
['could', 'better', 'hosts', 'missed', 'two', 'kickable', 'penalties', 'ireland', 'drew', 'level', 'ronan', 'penalty', 'midway', 'first', 'half']  
['italians', 'driving', 'heart', 'irish', 'defence', 'first', 'quarter', 'irish', 'pack', 'struggled', 'secure', 'ball', 'talented', 'backs']  
['finally', 'mark', 'promptly', 'created', 'sparkling', 'try', 'murphy']  
['ireland', 'captain', 'ran', 'dummy', 'scissors', 'made', 'magical', 'outside', 'break', 'drawing', 'putting', 'diving', 'murphy', 'corner']  
['missed', 'conversion', 'visitors', 'found', 'trailing']  
['roland', 'de', 'marigny', 'took', 'kicking', 'duties', 'italy', 'hapless', 'orquera', 'landed', 'penalty', 'either', 'side', 'break', 'edge', 'italy', 'lea']  
['ireland', 'player', 'offering', 'real', 'threat', 'break', 'set', 'second', 'try', 'visitors']  
['shane', 'horgan', 'threw', 'overhead', 'pass', 'forced', 'touch', 'stringer', 'scooted', 'landing', 'tricky', 'conversion']  
['penalty', 'apiece', 'saw', 'ireland', 'leading', 'game', 'entered', 'final', 'quarter', 'lucky', 'survive', 'italy', 'launched', 'series', 'attacks']  
['winger', 'nitoglia', 'dropped', 'ball', 'reached', 'line', 'italy', 'nearly', 'rumbled', 'driving', 'maul']  
['penalty', 'put', 'ireland', 'converted', 'try', 'ahead', 'made', 'game', 'safe', 'hickie', 'latched', 'onto', 'inside', 'pass', 'murphy', 'crossed', 'conve']  
['limped', 'late', 'joining', 'centre', 'partner', 'gordon', 'sidelines', 'final', 'word', 'went', 'italy']  
['prop', 'martin', 'castrogiovanni', 'powered', 'try', 'fitting', 'reward', 'italian', 'pack', 'kept', 'irish', 'pressure', 'throughout']  
['de', 'marigny', 'mi', 'bergamasco', 'canale', 'masi', 'nitoglia', 'orquera', 'troncon', 'lo', 'cicero', 'ongaro', 'castrogiovanni', 'dellape', 'bortolami',  
['perugini', 'intoppa', 'del', 'fava', 'dal', 'maso', 'griffen', 'pozzebon', 'robertson']  
['murphy', 'horgan', 'hickie', 'stringer', 'corrigan', 'byrne', 'hayes', 'easterby', 'leamy', 'foley']  
['sheahan', 'horan', 'miller', 'g', 'easterby', 'humphreys', 'dempsey']  
['p', 'new', 'zealand', 'british', 'irish', 'lions', 'coach', 'clive', 'woodward', 'says', 'unlikely', 'select', 'players', 'involved', 'next', 'year', 'rbs']  
['world', 'cup', 'winners', 'lawrence', 'dallaglio', 'neil', 'back', 'martin', 'johnson', 'thought', 'frame', 'next', 'summer', 'tour', 'new', 'zealand']  
['think', 'ever', 'say', 'never', 'said', 'woodward']  
['would', 'compulsive', 'reason', 'pick', 'player', 'available', 'international', 'rugby']  
['dallaglio', 'back', 'johnson', 'retired', 'international', 'rugby', 'last', '12', 'months', 'continue', 'star', 'club', 'sides']  
['woodward', 'added', 'key', 'thing', 'want', 'stress', 'intend', 'use', 'six', 'nations', 'players', 'available', 'international', 'rugby', 'key', 'benchmar']  
['job', 'along', 'senior', 'representatives', 'make', 'sure', 'pick', 'strongest', 'possible', 'team']  
['playing', 'international', 'rugby', 'still', 'step', 'test', 'rugby']  
['definitely', 'disadvantage']  
['think', 'absolutely', 'critical', 'history', 'lions', 'got', 'take', 'players', 'playing', 'four', 'countries']  
['woodward', 'also', 'revealed', 'race', 'captaincy', 'still', 'wide', 'open']  
['open', 'book', 'said']  
['outstanding', 'candidates', 'four', 'countries']
```

# Milestone 2: Building the Vocabulary

What is the resultant corpus\_count and corpus\_total\_words for all 511 articles?



```
[43] # What is the resultant corpus_count
print('Resultant corpus_count is:\n')
model.corpus_count

Resultant corpus_count is:
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel
8486

# What is the resultant corpus_total_words?
print('Resultant corpus_total_words are:\n')
model.corpus_total_words

Resultant corpus_total_words are:
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel
88917
```

**For all 511 articles** the average number of words per sentence for our dataset is **10.47 words per sentence**.

# Milestone 2: Training using all 511 articles

## ✓ Step 3.1: Training your model with the articles (excluding title sentences) assigned to your topic

✓  
32s



```
%%time  
model.train(tokenized_sentences, total_examples=model.corpus_count, epochs=30, report_delay=1)
```

```
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_cell`  
and should_run_async(code)  
CPU times: user 49.6 s, sys: 54.1 ms, total: 49.6 s  
Wall time: 31.8 s  
(1813284, 2667510)
```

# Top 3 sentences picked for a title using word2vec model trained for all 511 articles:

Title used:

```
and should_run_async(code)
['claxton', 'hunting', 'first', 'major', 'medal']
```

Top 3 sentences obtained:

```
WARNING:gensim.models.keyedvectors:destructive init_sims(replace=True) deprecated & no longer required for space-efficiency
WARNING:gensim.models.word2vec:Effective 'alpha' higher than previous training cycles
Title used for comparison: Claxton hunting first major medal
Sentence: british hurdler sarah claxton is confident she can win her first major medal at next month's european indoor championships in madrid.

Sentence: northern ireland man james mcilroy is confident he can win his first major title at this weekend's spar european indoor championships in madrid.

Sentence: and at last week's birmingham grand prix, claxton left european medal favourite russian irina shevchenko trailing in sixth spot.
```

# Part B: Milestone 2: Word2Vec

- We have done the summarization of the considered articles by taking title of one article as reference sentence.
- The reference sentence was loaded into the model and the model was trained again.
- Using Word2Vec model we have deduced the top 3 sentences based on the similarity scores with respect to the reference sentence.
- **model = Word2Vec(workers=4,min\_count=1, window=10, vector\_size=200, sg=1, sample=1e-3, negative= 10, alpha=0.025, min\_alpha=0.0001)**

# Summary Matching - before Hyper parameter tuning (Word2Vec Output)

Title used for comparison: Lions blow to World Cup stars

Corresponding Sentence: ['world', 'cup', 'winners', 'lawrence', 'dallaglio', 'neil', 'back', 'martin', 'johnson', 'thought', 'frame', 'next', 'summer', 'tour', 'new', 'zealand']

Corresponding Sentence: ['edu', 'brother', 'representative', 'amadeo', 'fensao', 'previously', 'said', 'arsenal', 'current', 'offer', 'midfielder', 'well', 'short', 'seeking']

Corresponding Sentence: ['told', 'back', 'playing', 'march', 'plenty', 'time', 'prove', 'fitness', 'lions', 'players', 'like', 'richard', 'hill', 'boat']

# Actual Summary

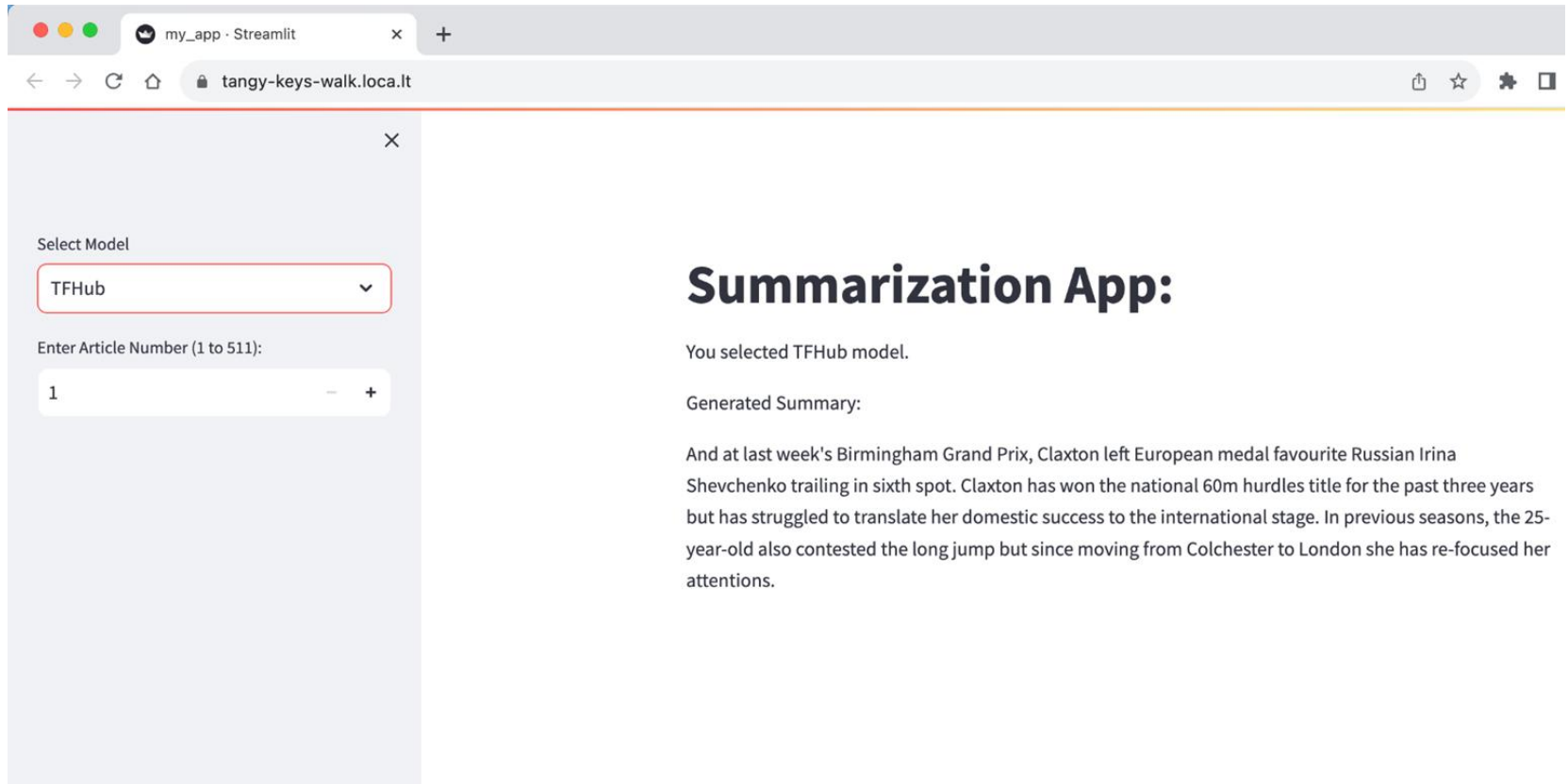
But Woodward added: "The key thing that I want to stress is that I intend to use the Six Nations and the players who are available to international rugby as the key benchmark."But I would have to have a compulsive reason to pick any player who is not available to international rugby."Every player has to be looked at on their own merits and Simon Taylor is an outstanding player and I have no doubts that if he gets back to full fitness he will be on the trip."I think it's absolutely critical and with the history of the Lions we have got to take players playing for the four countries."British and Irish Lions coach Clive Woodward says he is unlikely to select any players not involved in next year's RBS Six Nations Championship.As a result, Woodward stressed his Lions group might not be dominated by players from England and Ireland and held out hope for the struggling Scots."I am told he should be back playing by March and he has plenty of time to prove his fitness for the Lions - and there are other players like Richard Hill in the same boat."If you are not playing international rugby then it's still a step up to Test rugby.And following the All Blacks' impressive displays in Europe in recent weeks, including a 45-6 humiliation of France, Woodward believes the three-test series in New Zealand will provide the ultimate rugby challenge.Dallaglio, Back and Johnson have all retired from international rugby over the last 12 months but continue to star for their club sides.

# Part B: Milestone 2: Word2Vec

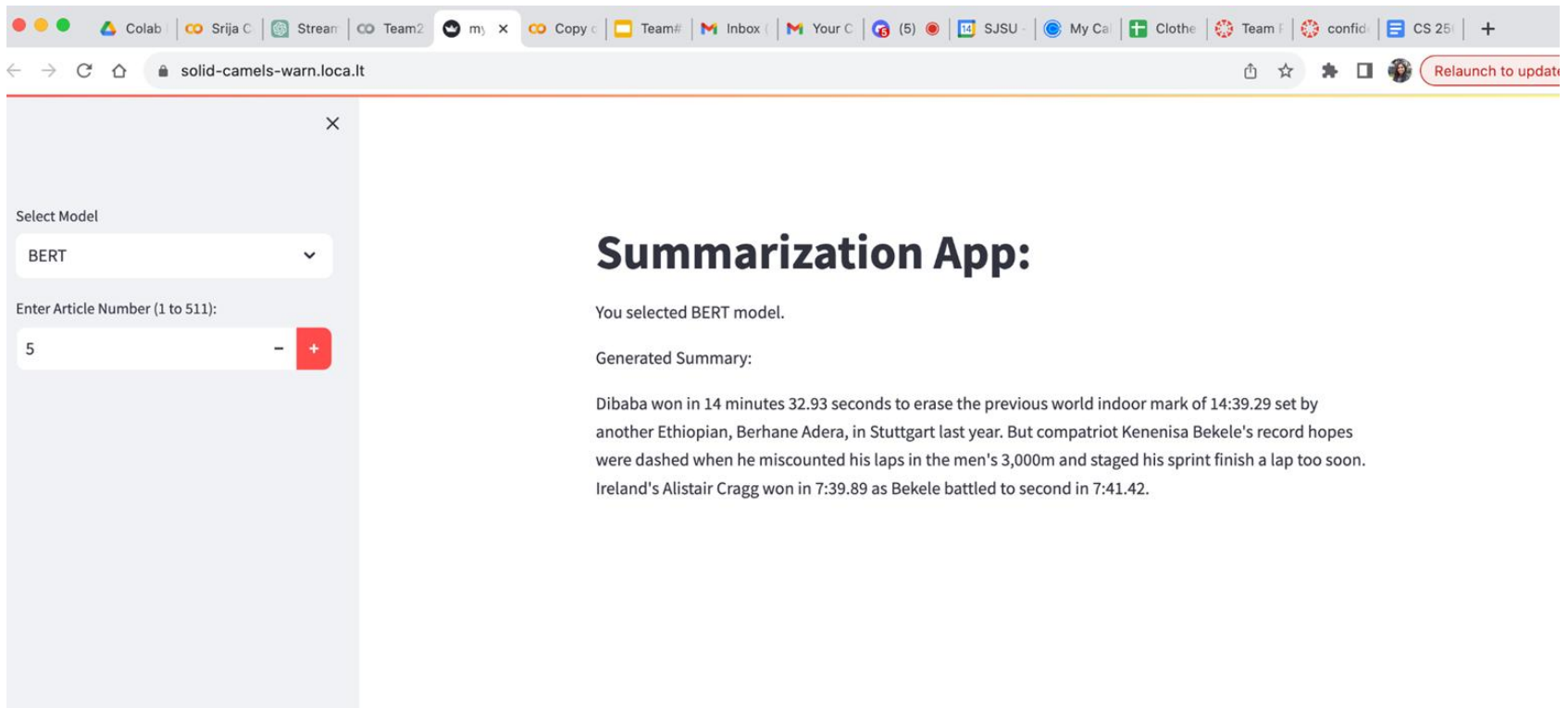
- We have done Hyper parameter tuning and tested the model for better results by taking the values of the parameters in the following range
- workers - [2,4,6,8]
- min-count - [0,1]
- vector\_size - [100, 200, 300, 400]
- negative sample size - [5,10,15]
- alpha - [0.01 - 0.04]
- min\_alpha - [0.00001,0.0001,0.001,0.1]
- After tuning hyper parameter in the above ranges we have found the following hyper parameters value to be more efficient.



# UI(Streamlit) ->TFHUB



# UI(Streamlit) ->BERT



**Thank You!**