UNIVERSITÉ CÔTE D'AZUR

MASTER INFORMATIQUE

News Articles Classification & Recommandation

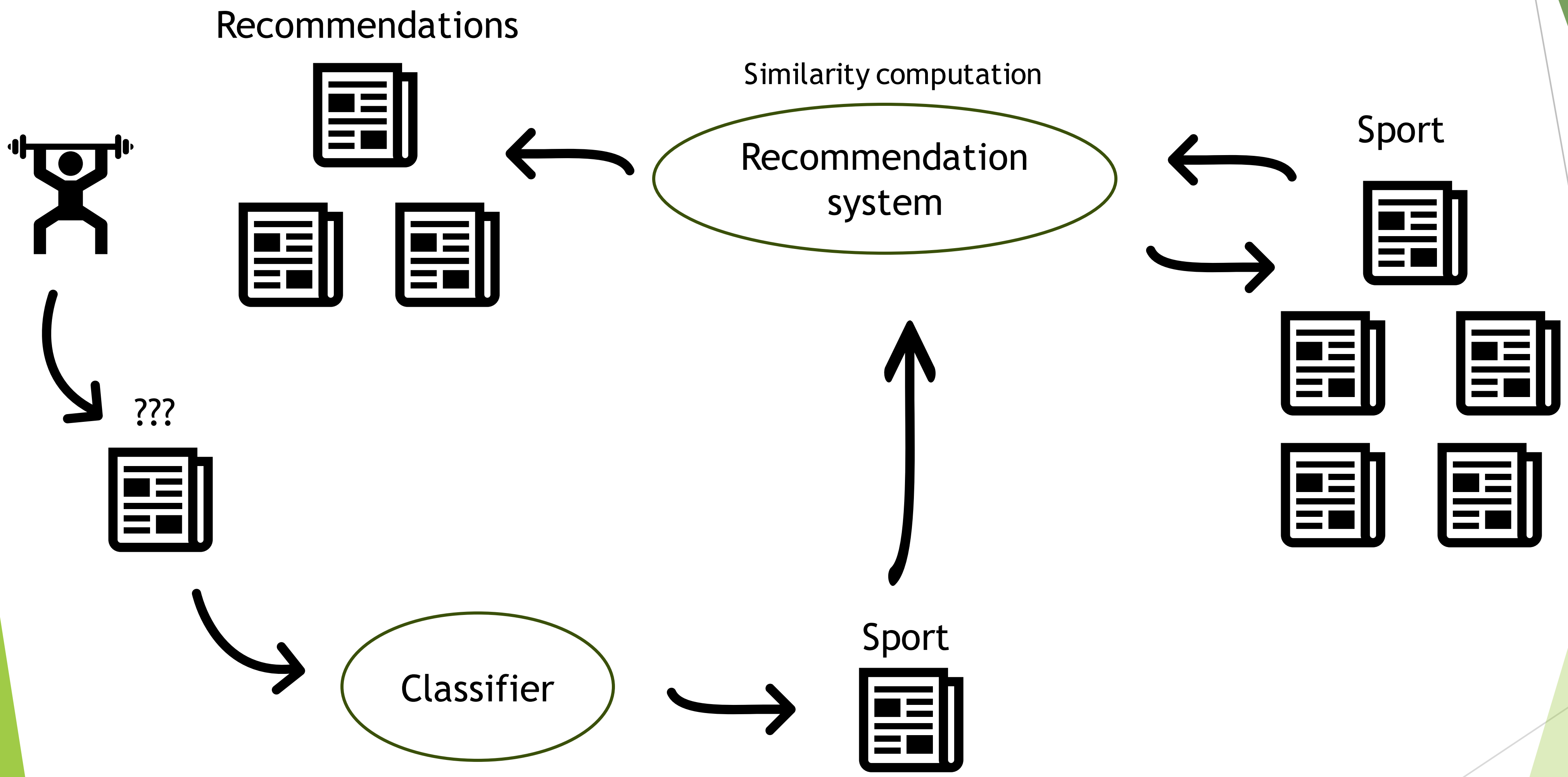
Amine HADDOU

Marouan BOULLI

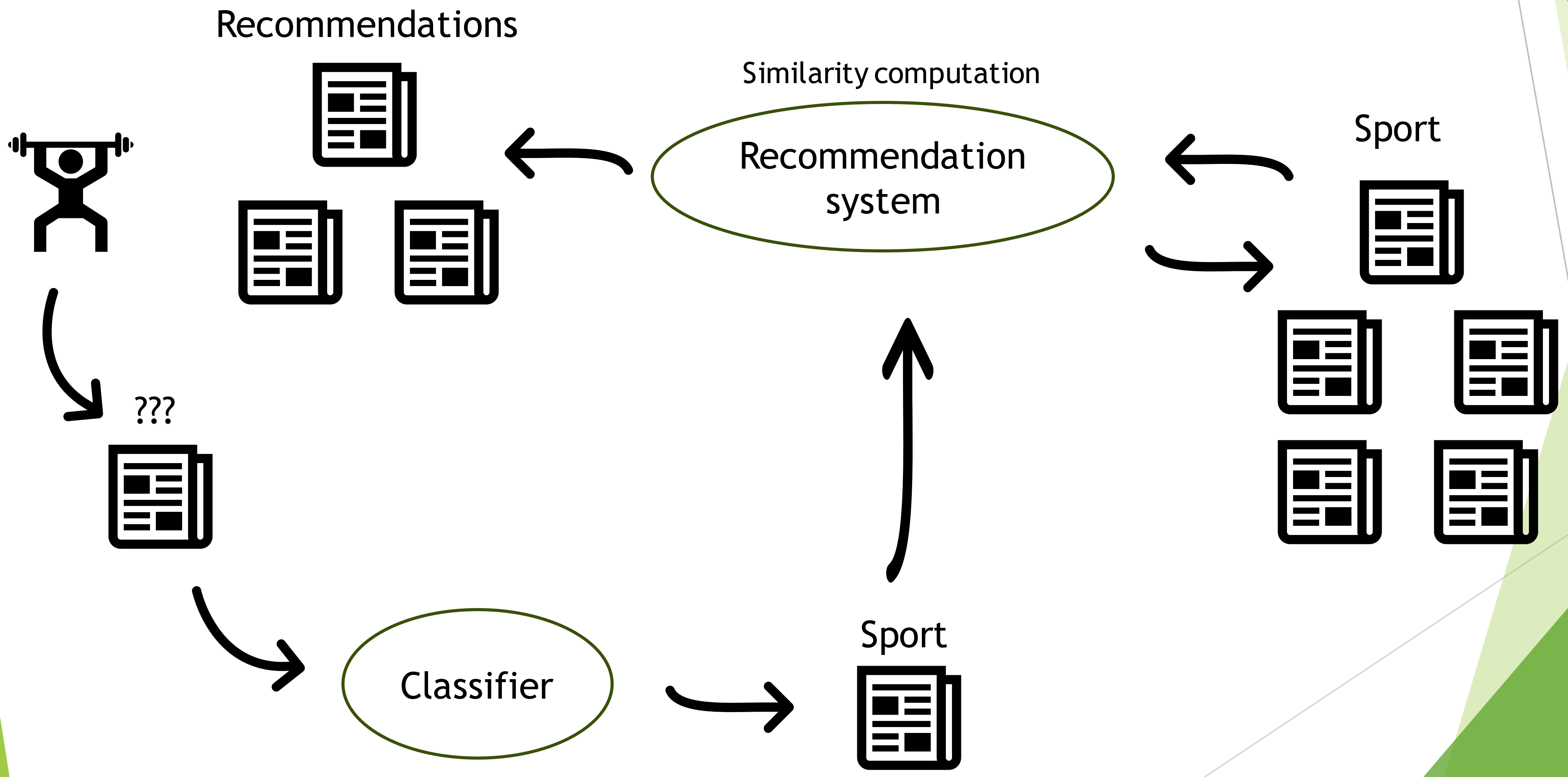
Tatia – 1st semester – 25/01/2024

Content

- Context
  - Purpose
  - Dataset
- Classification task
  - Preprocessing
  - Vectorization
  - Models training and evaluation
- Recommendation system
  - Articles similarity
  - Evaluations
- Conclusion


Recommendations
Similarity computation
Recommendation system
Sport
???
Classifier
Sport

Purpose


Recommendations
Similarity computation
Recommendation system
Sport
???
Classifier
Sport

Dataset

BBC News Classsification dataset
(https://www.kaggle.com/competitions/learn-ai-bbc)

Train 1490 labelled articles

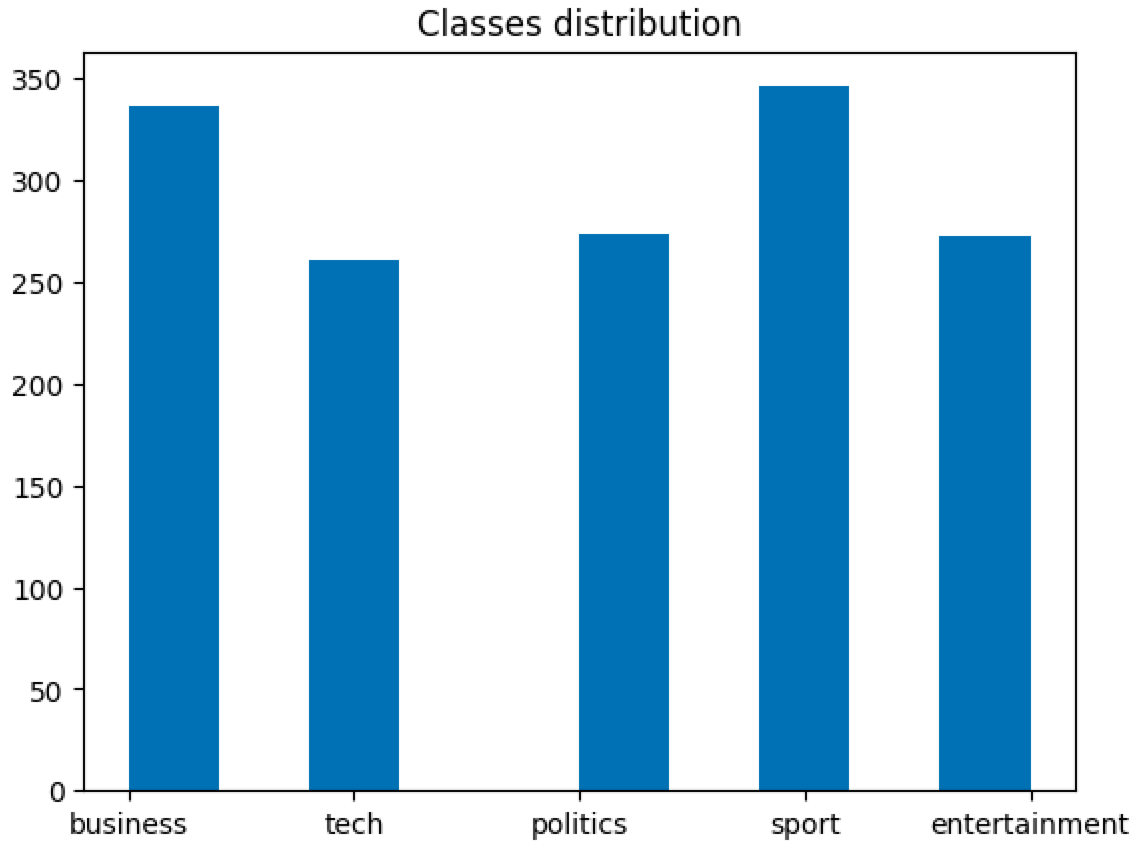
Test 735 non labelled articles

Business **Politics**

Sport **Tech**

Entertainment

Dataset


Classes distribution
350
300
250
200
150
100
50
0
business
tech
politics
sport
entertainment

Classification task

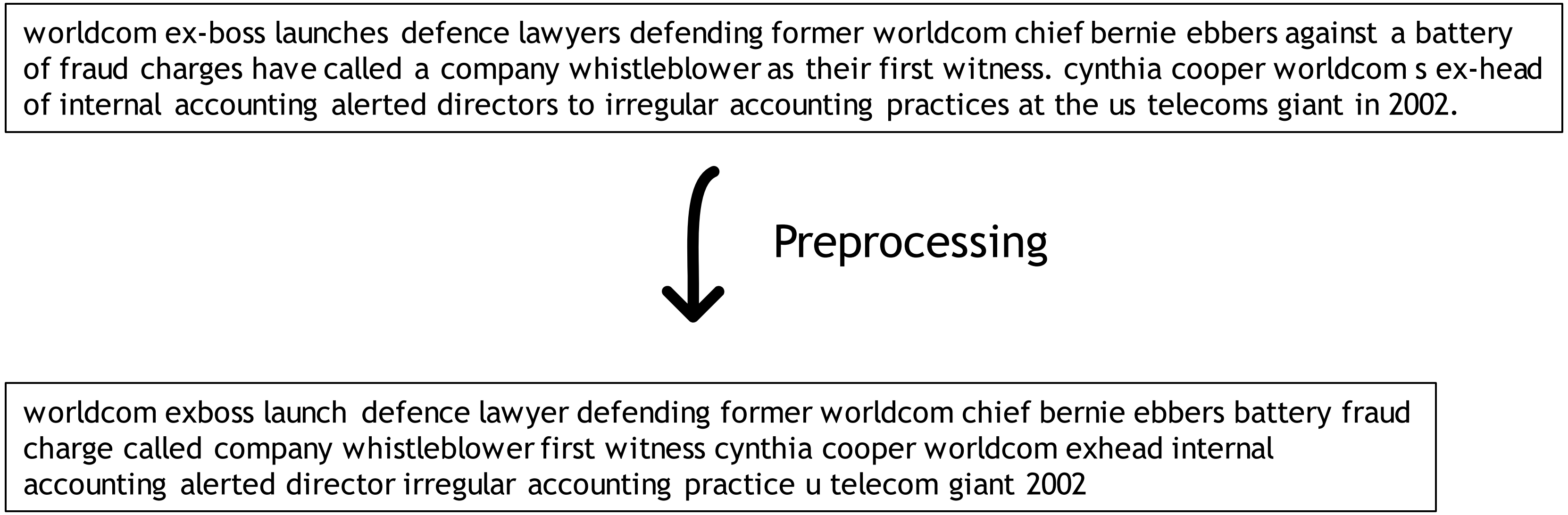
Preprocessing

Remove stop words

Remove punctuation

Lemmatization

Preprocessing



worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a battery of fraud charges have called a company whistleblower as their first witness. cynthia cooper worldcom s ex-head of internal accounting alerted directors to irregular accounting practices at the us telecoms giant in 2002.

worldcom exboss launch defence lawyer defending former worldcom chief bernie ebbers battery fraud charge called company whistleblower first witness cynthia cooper worldcom exhead internal accounting alerted director irregular accounting practice u telecom giant 2002

Vectorization

Count Vectorizer + TFIDF transformer

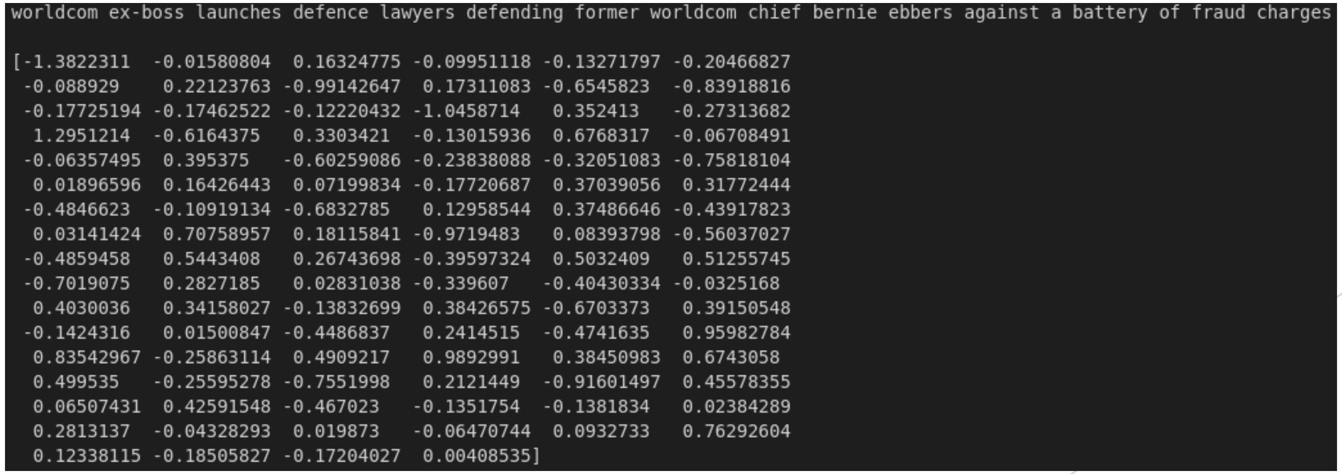
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Doc2Vec



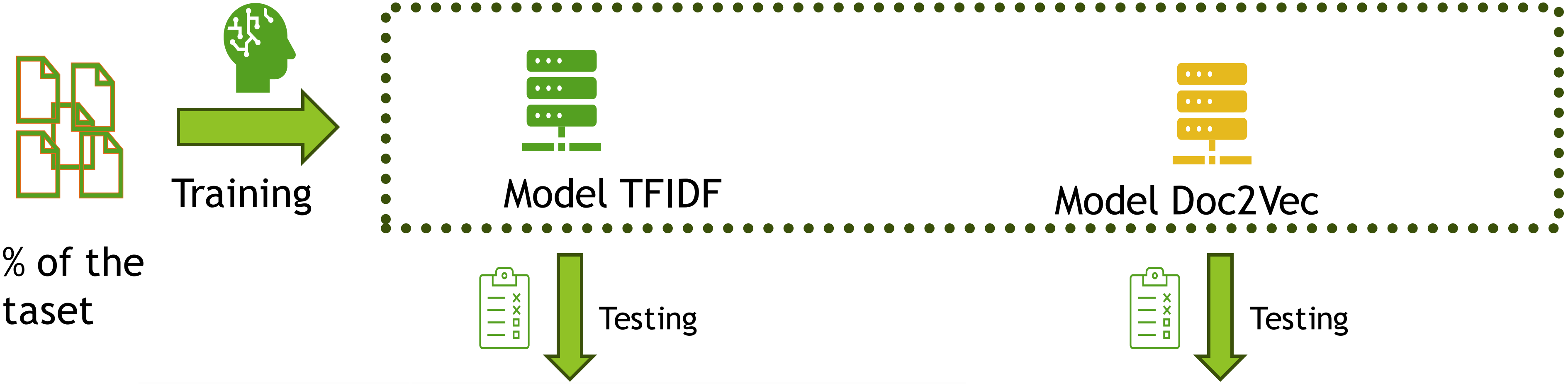
```
worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a battery of fraud charges

[-1.3822311  -0.01580804  0.16324775 -0.09951118 -0.13271797 -0.20466827
 -0.088929    0.22123763 -0.99142647  0.17311083 -0.6545823  -0.83918816
 -0.17725194 -0.17462522 -0.12220432 -1.0458714   0.352413   -0.27313682
  1.2951214  -0.6164375   0.3303421  -0.13015936  0.6768317  -0.06708491
 -0.06357495  0.395375   -0.60259086 -0.23838088 -0.32051083 -0.75818104
  0.01896596  0.16426443  0.07199834 -0.17720687  0.37039056  0.31772444
 -0.4846623  -0.10919134 -0.6832785   0.12958544  0.37486646 -0.43917823
  0.03141424  0.70758957  0.18115841 -0.9719483   0.08393798 -0.56037027
 -0.4859458   0.5443408   0.26743698 -0.39597324  0.5032409   0.51255745
 -0.7019075   0.2827185   0.02831038 -0.339607   -0.40430334 -0.0325168
  0.4030036   0.34158027 -0.13832699  0.38426575 -0.6703373   0.39150548
 -0.1424316   0.01500847 -0.4486837   0.2414515  -0.4741635   0.95982784
  0.83542967 -0.25863114  0.4909217   0.9892991   0.38450983  0.6743058
  0.499535   -0.25595278 -0.7551998   0.2121449  -0.91601497  0.45578355
  0.06507431  0.42591548 -0.467023   -0.1351754  -0.1381834   0.02384289
  0.2813137  -0.04328293  0.019873   -0.06470744  0.0932733   0.76292604
  0.12338115 -0.18505827 -0.17204027  0.00408535]
```

Models Training and Evaluation

- Support Vector Machine
- K-neighbors Classifier

Support Vector Machine

- **Support vector machines (SVMs)** are a set of supervised learning methods used for classification, regression and outliers detection.



70% of the dataset

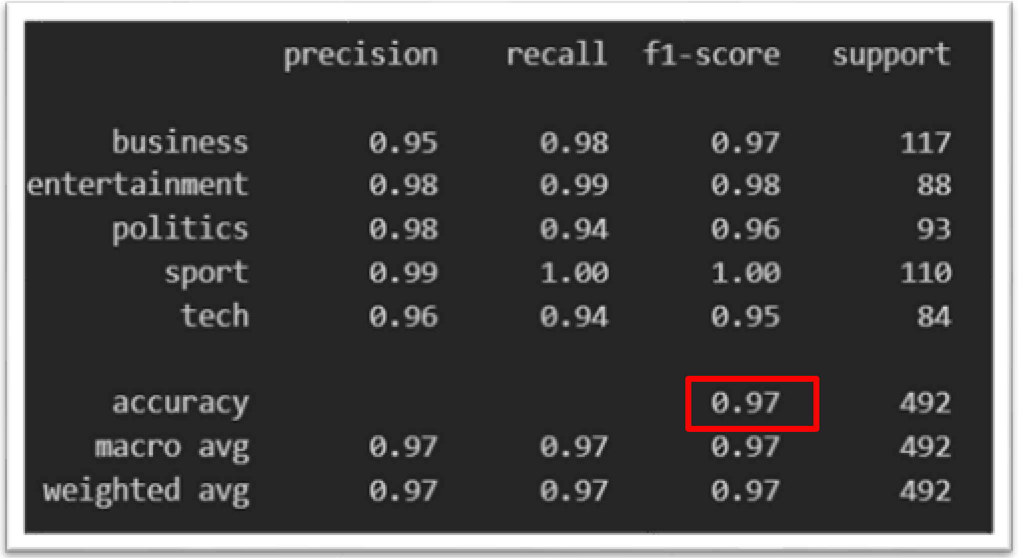
Training

Model TFIDF

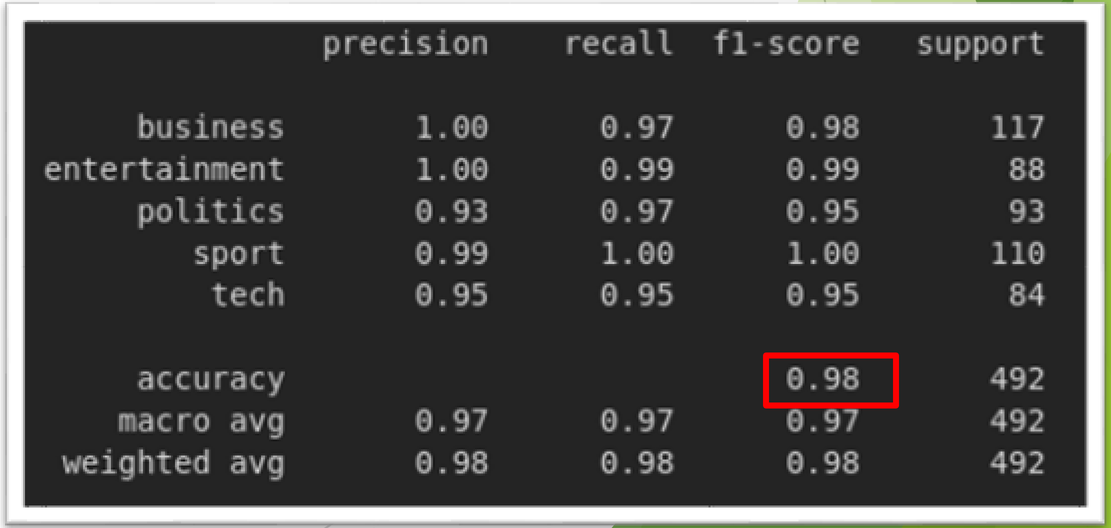
Model Doc2Vec

Testing

Testing



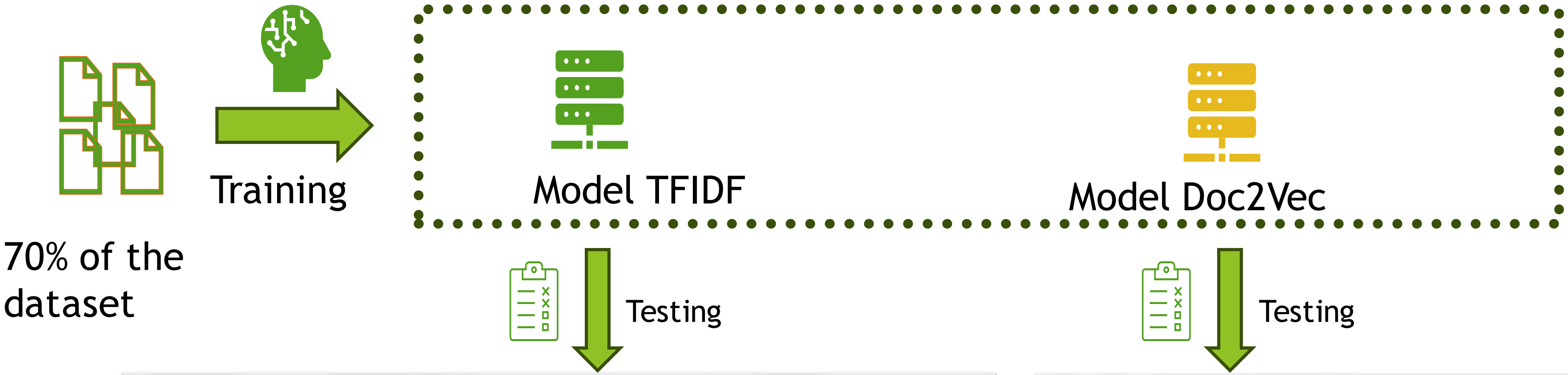
```
              precision    recall  f1-score   support

    business       0.95      0.98      0.97       117
entertainment      0.98      0.99      0.98        88
    politics       0.98      0.94      0.96        93
       sport       0.99      1.00      1.00       110
        tech       0.96      0.94      0.95        84

    accuracy                           0.97       492
   macro avg       0.97      0.97      0.97       492
weighted avg       0.97      0.97      0.97       492
```



```
              precision    recall  f1-score   support

    business       1.00      0.97      0.98       117
entertainment      1.00      0.99      0.99        88
    politics       0.93      0.97      0.95        93
       sport       0.99      1.00      1.00       110
        tech       0.95      0.95      0.95        84

    accuracy                           0.98       492
   macro avg       0.97      0.97      0.97       492
weighted avg       0.98      0.98      0.98       492
```

K Neighbors Classifier

- Classifier implementing the k-nearest neighbors vote.



Training

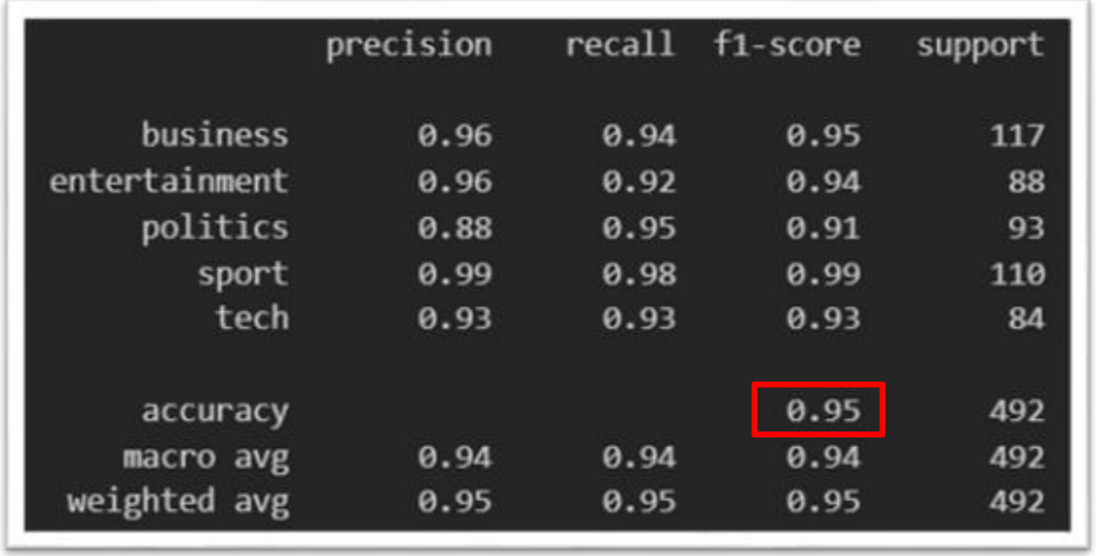
Model TFIDF

Model Doc2Vec

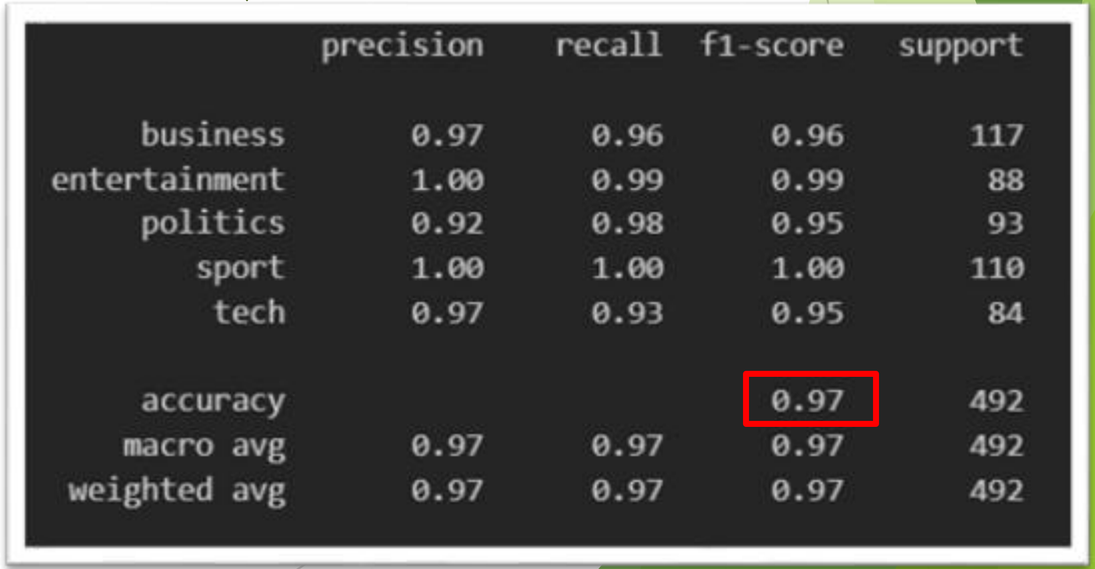
70% of the dataset

Testing

Testing



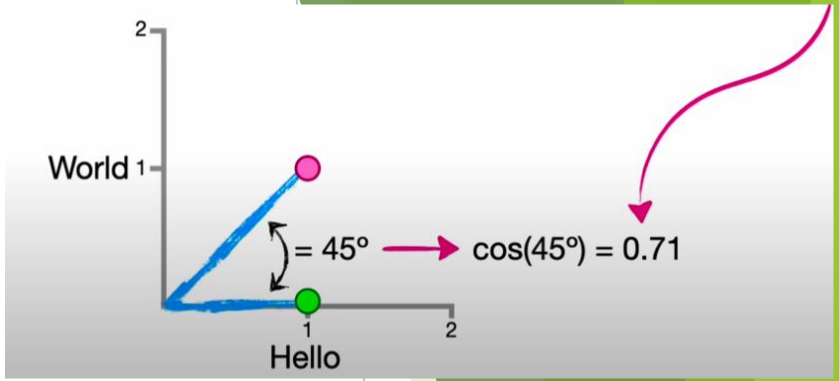
```
               precision    recall  f1-score   support

     business       0.96      0.94      0.95       117
entertainment       0.96      0.92      0.94        88
     politics       0.88      0.95      0.91        93
        sport       0.99      0.98      0.99       110
         tech       0.93      0.93      0.93        84

     accuracy                           0.95       492
    macro avg       0.94      0.94      0.94       492
 weighted avg       0.95      0.95      0.95       492
```

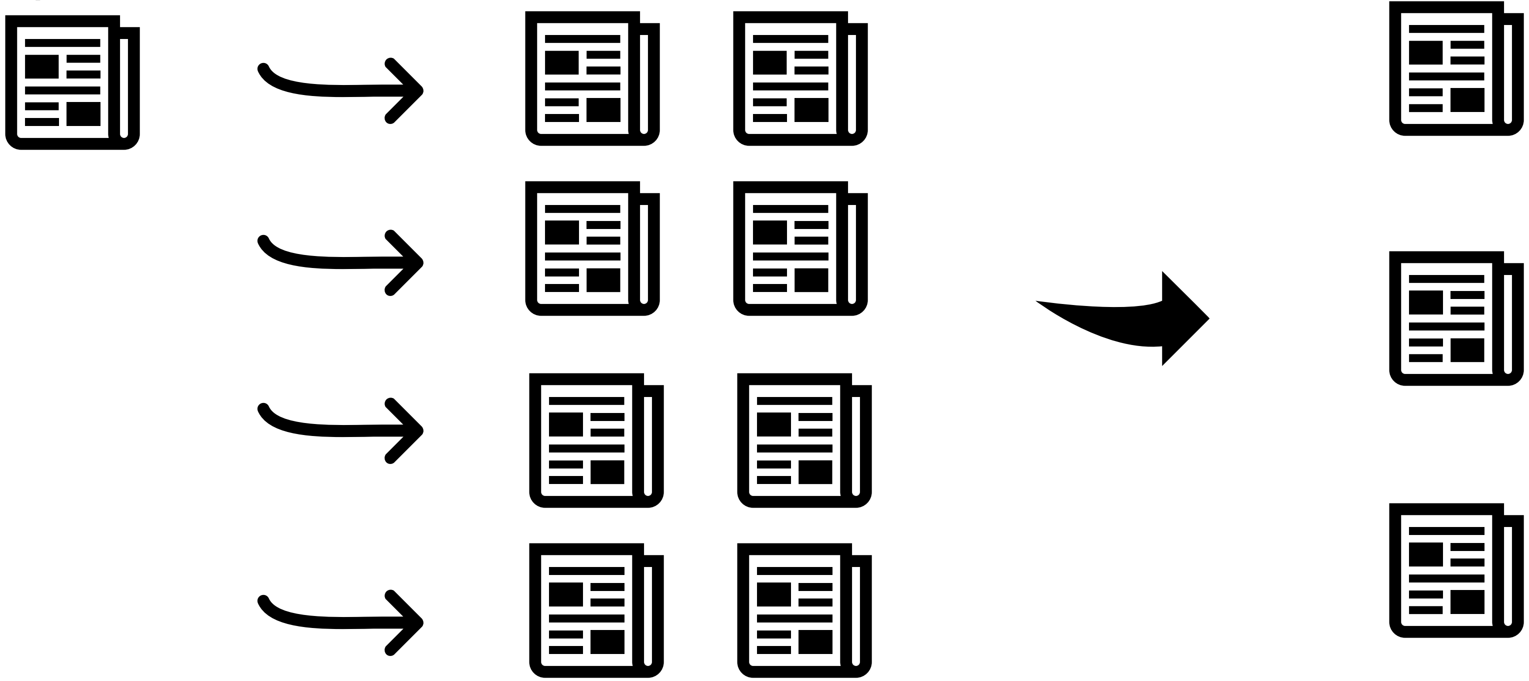


```
               precision    recall  f1-score   support

     business       0.97      0.96      0.96       117
entertainment       1.00      0.99      0.99        88
     politics       0.92      0.98      0.95        93
        sport       1.00      1.00      1.00       110
         tech       0.97      0.93      0.95        84

     accuracy                           0.97       492
    macro avg       0.97      0.97      0.97       492
 weighted avg       0.97      0.97      0.97       492
```

Recommendation task

Articles similarity





Sport

Sport

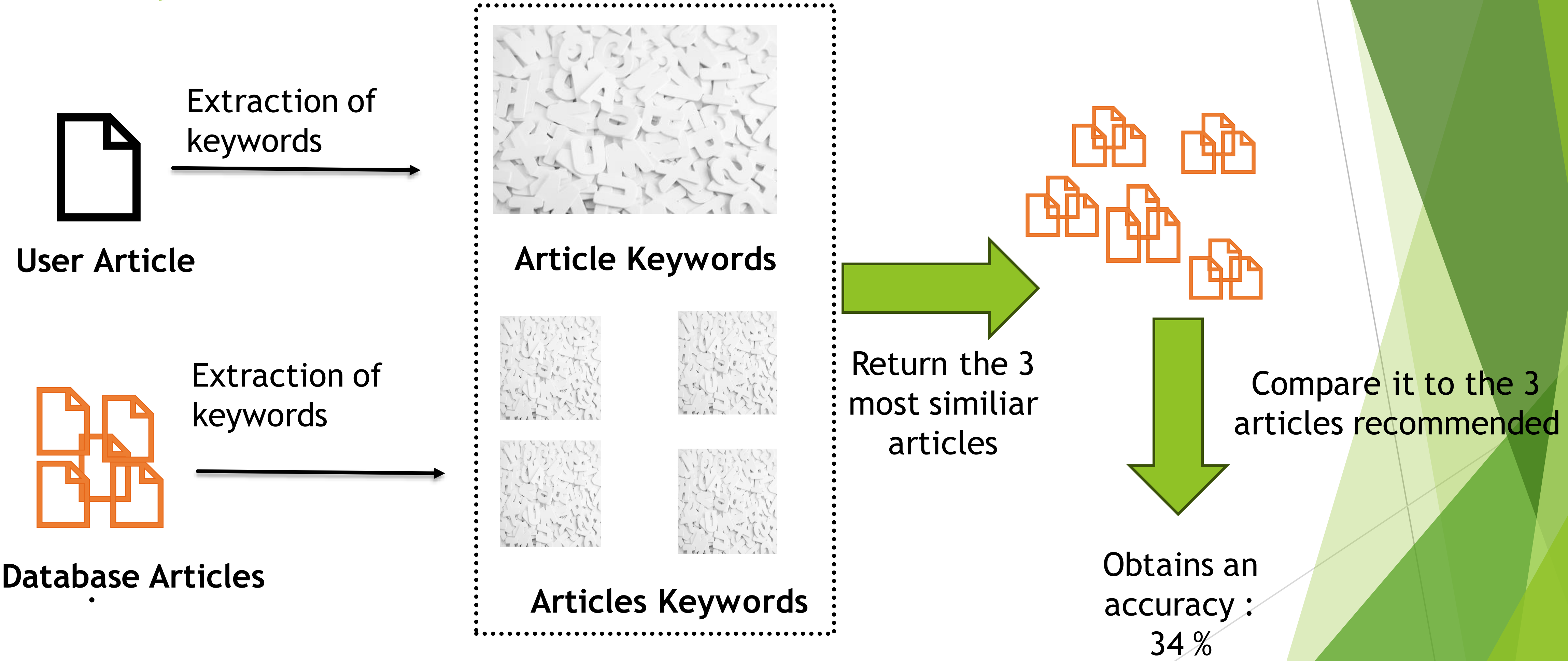
Top 3 cosine scores

Cosine similarity

Evaluations

- ✓ Keywords extraction
- ✓ LLM Evaluation

Keywords extraction



User Article

Extraction of keywords

Article Keywords

Database Articles

Extraction of keywords

Articles Keywords

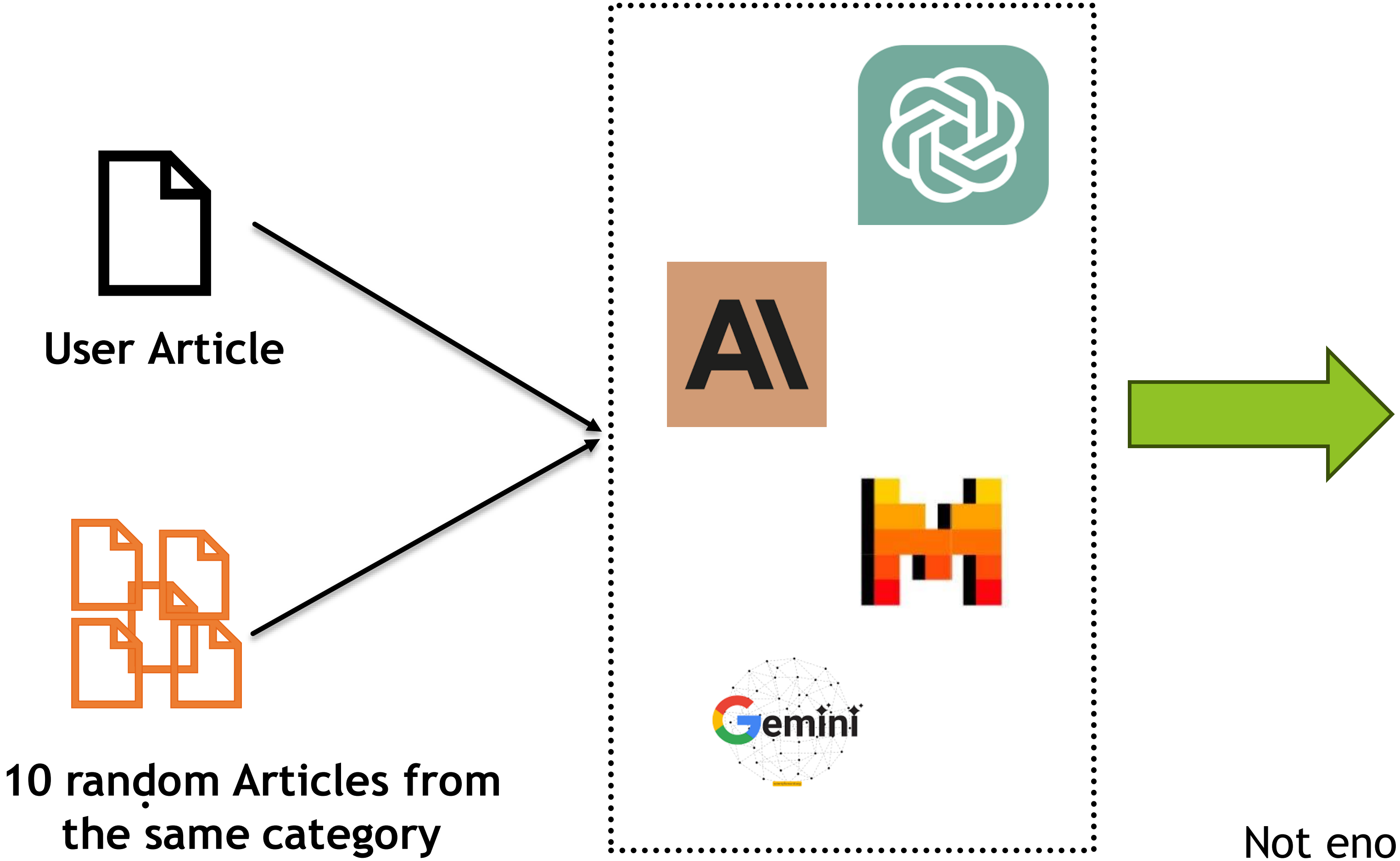
Return the 3 most similiar articles

Compare it to the 3 articles recommended

Obtains an accuracy : 34 %

For every article, we compute the number of **common keywords** with the others

LLM Evaluation



User Article

10 random Articles from the same category

On 10 articles :

- 2 were redundants
- 1 was "unique"

- On the 3 recommended articles, only one appeared in the LLMs' classements

Not enough to conclude but it gives an inidcation on the quality of the recommandation

Conclusion

- Classification + recommendation
- High classification accuracy
- Vectorization method is important
- Automated evaluation of the recommendation system is not easy
- Improvements:
  - Provide the LLMs with the entire dataset to improve the reliability of evaluation
  - Consider performing an end-user evaluation
  - Test the recommendation without classifying the article

Thank you!