



# Suggester Evaluation Level

24.08.2023



# Suggester Evaluation Level



# Part Contents

## 1 Suggester Evaluation Level



# Section Contents

## 1 Suggester Evaluation Level

### Introduction

An Upper Bound for Cosine Distance

Changes to the UI

Conclusions and Next Steps



# Introduction

- ▶ So far the model suggester simply identifies whether a document was to considered relevant or not based on the evaluation criteria;
- ▶ Instead, we would need a way of identifying suggestion levels, such as RELEVANT, POTENTIALLY RELEVANT or NOT RELEVANT;
- ▶ We need then to modify the suggestion algorithm accordingly.



# Section Contents

- 1 Suggester Evaluation Level**
  - Introduction
  - An Upper Bound for Cosine Distance**
  - Changes to the UI
  - Conclusions and Next Steps



# The Cosine Distance

- ▶ The criteria to determine whether a document is relevant or not is based on the computation of the *cosine distance* between its vectorized form and the vectorized form of the relevant documents we have in the training set;
- ▶ The minimum distance is taken and compared with a threshold value. If it is lower than the threshold value the document is considered relevant, otherwise not.



## An Upper Bound for Cosine Distance

What we can do is trying to find a reasonable upper limit for the distance in such a way that:

- ▶ if the distance is below the threshold, the document is **RELEVANT**;
- ▶ if the distance is above the threshold but below the upper limit, the document is **POTENTIALLY RELEVANT**;
- ▶ if the distance is above the upper limit, the document is **NOT RELEVANT**.

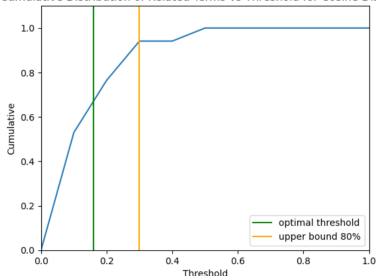




## An Upper Bound for Cosine Distance

- ▶ We considered the cumulative distribution of the distances for the test set compared to the train set;
- ▶ We took as upper limit the value of the distance which corresponds to 80% of the truly relevant document properly identified.

Cumulative Distribution of Related Terms vs Threshold for Cosine Distance





# Section Contents

- 1 Suggester Evaluation Level**
  - Introduction
  - An Upper Bound for Cosine Distance
  - Changes to the UI**
  - Conclusions and Next Steps



## Evaluation Mechanism

- ▶ When evaluating a model now, the relevant documents are displayed in red, the potentially relevant in orange and the not relevant in black;
- ▶ The user still has the possibility to adjust all these predictions and retraining the model;
- ▶ All documents saved as relevant or potentially relevant will end up in the new training set as relevant (this might be adjusted at a later point).

| Feature Name               | Evaluated Docs |                      |
|----------------------------|----------------|----------------------|
|                            | Document       | Relevance Level      |
| Address::city              | city           | RELEVANT             |
| Address::context           | context        | POTENTIALLY_RELEVANT |
| <a href="#">Save Model</a> |                |                      |



## Visualization View

- ▶ When a model is loaded the relevant features are displayed in red, the potentially relevant in orange and the not relevant in black.

|                              |                     |             |             |
|------------------------------|---------------------|-------------|-------------|
| Select the criteria:<br>GDPR |                     |             |             |
| Address                      | EClassifier         |             |             |
|                              | EStructuralFeatures |             |             |
|                              | Name                | EType       | Cardinality |
|                              | city                | EString     | 0..1        |
|                              | context             | ContextType | 0..1        |
|                              | id                  | EString     | 1..1        |
|                              | state               | EString     | 0..1        |
|                              | street              | EString     | 0..1        |
|                              | zip                 | EString     | 0..1        |



# Section Contents

## 1 Suggester Evaluation Level

Introduction

An Upper Bound for Cosine Distance

Changes to the UI

Conclusions and Next Steps



## Conclusions and Next Steps

- ▶ We found a mechanism to determine which documents are relevant and to assign them a certain level of relevance;
- ▶ We might want to think of a better way to retrain the model, which immediately takes into account the difference between relevant and potentially relevant;
- ▶ We have to implement the suggestion mechanism for all EClassifier (currently we are only doing that for EClass).



# Conclusion



## Useful Links

### OSGi Working Group

Working Group: [www.osgi.org](http://www.osgi.org)

WG Blog: [www.osgi.org/blog](http://www.osgi.org/blog)

Twitter: [@osgiwg](https://twitter.com/osgiwg)

Bndtools: [bndtools.org](http://bndtools.org)

### Data In Motion

Web: [www.datainmotion.com](http://www.datainmotion.com)

Blog: [datainmotion.com/blog](http://datainmotion.com/blog)

Twitter: [@motion\\_data](https://twitter.com/motion_data)

### Jürgen Albert

Email: [j.albert@data-in-motion.biz](mailto:j.albert@data-in-motion.biz)

### Mark Hoffmann

Email:  
[m.hoffmann@data-in-motion.biz](mailto:m.hoffmann@data-in-motion.biz)