Information Extraction from Video Webinar Recordings

Machine Learning Project
University of Tartu
2021
Team

Ramil Huseynov  
Soft. Eng. MSc

Ashraf Abbasov  
Comp. Sci. MSc

Dariya Nagashibayeva  
Comp. Sci. MSc

Sarp Aktug  
Soft. Eng. MSc

Alexander Tkachenko  
Product Owner (Machine Learning Engineer) at Snackable.ai
Problem

Video content is not easy to search using traditional text-based searching.

Tags are used to overcome this.

Yet, they are done manually, which takes a lot of time and human effort.
Solution
Our Approach

- Extract slides from the webinar video
- Extract text from the slides
- Generate candidates and their features from the text
- Classify candidates as keywords
Extract slides from the webinar video

OpenCV

Hashing
Extract text from the slides

Tesseract OCR to read text with features (location, size, capitalization)
Generate candidates and their features from the text

Spacy to generate candidates list
Classifying the candidates - Training model

- Feature engineering
Feature Engineering

- Open CV from video
  - Occurrence timestamps (first, last)
- Tesseract OCR from slides
  - Location
  - Frequency
  - Font size
  - Capitalization
- Spacy NLP from text
  - Part of speech
  - Label (Organization, Person etc.)
- Keybert score
- Words count
Classifying the candidates - Training model

- Feature engineering
- Feature selection
  - ANOVA f-classif
Classifying the candidates - Training model

- Feature engineering
- Feature selection
  - ANOVA f-classif
- Dimensionality reduction
  - PCA

![Diagram showing process flow from slides to keyphrases](image-url)
Classifying the candidates - Training model

- Feature engineering
- Feature selection
  - ANOVA f-classif
- Dimensionality reduction
  - PCA
- Model selection

![Diagram showing the process of classifying candidates](image)
## Model Selection

<table>
<thead>
<tr>
<th>Model</th>
<th>ROC_AUC_score</th>
<th>Highest f1_score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.78</td>
<td>0.34</td>
</tr>
<tr>
<td>GaussianNB</td>
<td>0.62</td>
<td>0.15</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>0.76</td>
<td>0.35</td>
</tr>
<tr>
<td>MLPClassifier</td>
<td>0.82</td>
<td>0.37</td>
</tr>
<tr>
<td>XGBClassifier</td>
<td>0.76</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Classifying the candidates - Training model

- Feature engineering
- Feature selection
  - ANOVA f-classif
- Dimensionality reduction
  - PCA
- Model selection
- Tuning the model
## Results

<table>
<thead>
<tr>
<th>True keyphrases</th>
<th>Generated by our model</th>
<th>Keybert’s top 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>['artificial intelligence', 'business', 'business process', 'data', 'data science central', 'innovation', 'leonardo', 'sap', 'sap leonardo', 'software', 'technologies', 'technology']</td>
<td>['artificial intelligence', 'data', 'data science central', 'design', 'dsc', 'dsc webinar', 'empathy', 'innovation', 'leonardo', 'october', 'process', 'project', 'sap', 'sap leonardo', 'saphana', 'tableau', 'technology']</td>
<td>['dsc webinar', 'data scientist', 'data science central', 'sap innovation centers', 'sap', 'dsc', 'global sap leonardo', 'global ceo', 'technology partners', 'global ceos', 'digital ecosystem', 'technology partner', 'global ceo outlook', 'sap data hub', 'sap labs']</td>
</tr>
</tbody>
</table>
Lessons

• Importance of
  ○ data collection
  ○ data engineering
  ○ data preprocessing

• Applied theories covered in the course
Q & A