Highlight Detection

Pavel Chizhov
Ivan Lastovko
Denys Krupovych
The Task
Audio/Text Summarization

[SNACKABLE]

Original TED talk: 20 minutes

AI-generated audiogram: 1 minute

We want the same for audio content
The Data

Generated Transcripts

Model

Highlights:
- Timecodes
- Scores
Our Solution

1. **Preprocessing**
   - Lancaster stemming and stopwords/punctuation filtering

2. **Vectorization**
   - Word2Vec
   - TF-IDF Weightings

3. **RNN**
   - Bidirectional GRU + Dense
   - Binary probabilistic output

4. **Join&Filter**
   - Joining sequential highlights and filtering the shortest
Results

Val nDCG@10 = 0.258, Test nDCG@10 = 0.228
(Baseline = 0.170, Human = 0.449)
Main obstacles

- **Lack of experience**
  - Unknown field and unusual format

- **High load**
  - Time flies faster than we thought

- **Unstable behaviour**
  - Varying scores on different iterations
Lessons Learned

01  Working with texts
    Basics of stemming and vectorization

02  Intro to RNNs
    Input embeddings
    Recurrent layers

03  New metric
    nDCG
    Different from NN metric

04  Teamwork
    Brainstorming
    Time-management
Q&A