**Industry mapping using scraped text**

Project in data collection lab

*Technion, 2022*

Snir Lugassy, Alon Shiri

Abstract

Text processing and normalization

* Filtering English text
  + Latin characters threshold
  + Country filtering
* Normalization pipeline
  + Remove HTML tags
  + Remove special characters
  + Remove digits
  + Remove redundant whitespace
  + Convert to lowercase
  + Word length filtering: only word with length between 3 and 14

Word embedding using tf-idf

Adding “related to” column using 1-NN

Map reduce