# Extraction of Skills and Benefits from Job Postings Descriptions

Shchapaniak Andrei

# Overview

## Job Description

| | |
|---|---|
| **Job title:** | Insert job title |
| **Location:** | Where is job located? Any travel etc? |
| **Terms:** | Perm/contract? Hours? Full/part time? |
| **Salary/rate:** | Include remuneration if possible |
| **Requirements:** | Any special requirements such as weekend work |

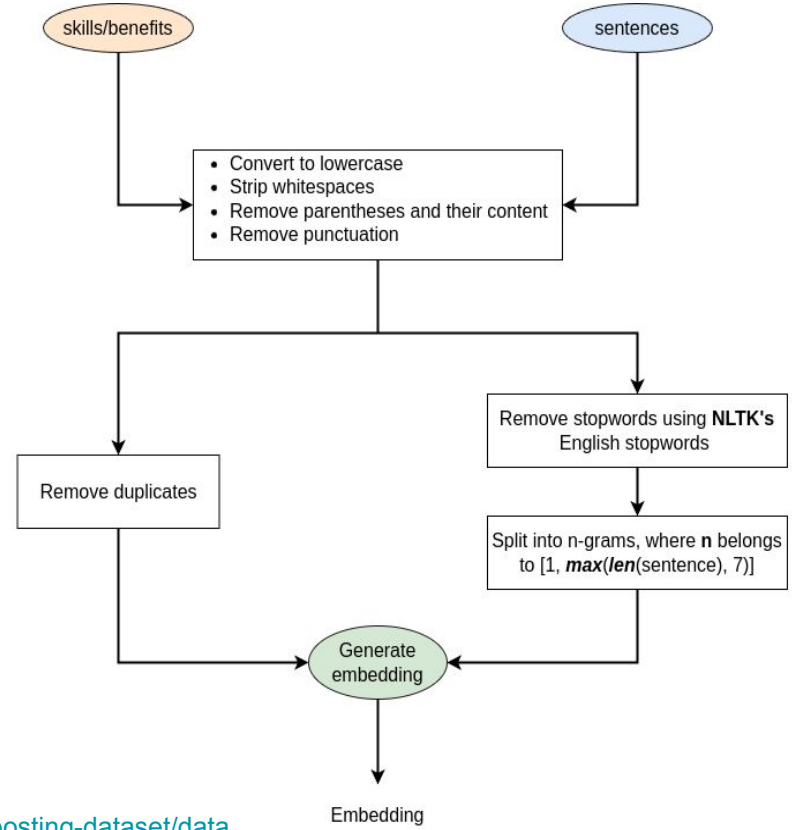
| | |
|---|---|
| **About us:** | Brief description of your organisation, such as what markets they operate in, products and services offered, mission statement, culture and values etc. |
| **Skills** | Experienced software engineer with 4+ years commercial experience in any programming language, e.g. Python.<br>Strong mastery of software architecture with a flair for writing extensible, maintainable code.<br>Knowledge of basic algorithms and data structures.<br>Solid level of spoken and written English.<br>Enthusiasm and a natural ability to work well with others and be a team -player.<br>Experience with technical leadership is a big plus. |
| **Benefits:** | • Medical Insurance<br>• 401k<br>• Transport Benefits<br>• Career Coaching<br>• Work Equipment |
| **Candidate requirements:** | • Bullet pointed list of skills, experience and qualifications successful candidates will need<br>• Be specific as possible, using numbers where possible |

**Skills**: [‘Python’, ‘algorithms’, ‘data structures’, ‘english’, ‘programming’, ‘leadership’]

**Benefits**: [‘medical insurance’, ‘401k’, ‘transport benefits’, ‘career coaching’, ‘work equipment’]

# Data & Preprocessing

- **Indeed Job Posting Dataset** - main dataset, contains job descriptions
- **Job Dataset** - contains features with well-known benefits
- **ESCO Skills Dataset** - contains huge set of well-known skills (> 20000)
- ChatGPT Generated Benefits (50 values)




Indeed Job Posting Dataset: https://www.kaggle.com/datasets/promptcloud/indeed-job-posting-dataset/data
Job Dataset: https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset
ESCO Skills Dataset: https://www.kaggle.com/datasets/thenoob69/esco-skills

# Model perfomance

| | Model | Accuracy | Recall | F1-score |
|---|---|---|---|---|
| **Skills** | MiniLM | **0.57** | **0.40** | **0.47** |
| | JobBert | 0.12 | 0.30 | 0.17 |
| | Roberta | 0.08 | 0.20 | 0.11 |
| | DistilBert | 0.11 | 0.40 | 0.17 |
| | | | | |
| **Benefits** | MiniLM | **1.00** | **0.57** | **0.73** |
| | JobBert | 0.67 | **0.57** | 0.62 |
| | Roberta | 0.44 | **0.57** | 0.50 |
| | DistilBert | 0.22 | **0.57** | 0.32 |

# Optimization roadmap

- **Embedding caching** - {1,2,3}-grams can repeat => utilize LRU caching for them
- **Efficient n-gram generation** - don’t add duplications, custom window slicing, no temporary list employing
- **Enhanced similarity calculations** - remove unnecessary type/device conversions, *torch.no_grad* to avoid intermediate results storage, early mask application with threshold

# Results & Future work

**15** random samples from dataset:

- **Original Extractor** - 4min 26s
- **Optimized Extractor** - 3min 41s (**17%** acceleration)

---

**Future possible improvements:**

- Reduce terminology variance (~dataset size, decrease #false_positive)
- Better text preprocessing
- Parallelization of skills/benefits processing