Log Book  
Pesquisa na Web

short line

# Objetivo

Log Book para os updates que vamos fazer ao longo das semanas no âmbito da disciplina de Pesquisa na Web.

**Professor:** João Magalhães **Tema:** Open-Domain search assistant

**Grupo:**

* Gonçalo Antunes
* João Diogo
* Samuel Viegas

# Semanas

## 25/03

Info: Iniciar e preparar o projecto

* Preparar o Colab
* Preparar os dados iniciais - conversas

Cenário:

* 20 conversas
* 8 perguntas em cada conversa
* 20 parágrafos candidatos a resposta possível a cada pergunta

**Passos possíveis:**

Linha de trabalho 1 (entity linking):

* Extrair entidades das perguntas e das passagens

Linha de trabalho 2 (rank das entidades):

* Construir grafo de entidades (ligação entre entidade é a co-ocorrência numa frase)
* Aplicar o PageRank no grafo de entidades
  + <https://scikit-learn.org/stable/auto_examples/applications/wikipedia_principal_eigenvector.html>
* Obter o PageRank de cada entidade
* Calcular a importância de cada passagem de acordo com as suas entidades
* Escolher as passagens mais importantes (possívelmente mais focadas nas mesmas entidades)

Linha de trabalho 3 (geração da resposta):

* Passar as passagens mais importantes ao BART para gerar a resposta única (biblioteca do huggingface)

Oc=

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | p1 | p2 | p3 | p4 |
| e1 (car) | 1 |  | 1 |  |
| e2 (sun) |  | 1 | 1 |  |
| e3 (beach) |  | 1 |  |  |
| e4 | 1 | 1 |  | 1 |
| e5 | 1 |  | 1 |  |

Graph = Oc \* Oc.T =

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | e1 | e2 | e3 | e4 | e5 |
| e1 | 2 | 1 | 0 |  |  |
| e2 | 1 | 2 | 1 |  |  |
| e3 | 0 | 1 | 1 |  |  |
| e4 |  |  |  |  |  |
| e5 |  |  |  |  |  |

## 01/04

Férias

## 08/04

## **Estrutura do Relatório**

Introduction

(o trabalho é sobre? 1) comparação de métodos, 2) um método com um dado input e output, 3) um sistema com dadas funcionalidades, 4) etc)

Algorithms

Implementation

Evaluation

Dataset description

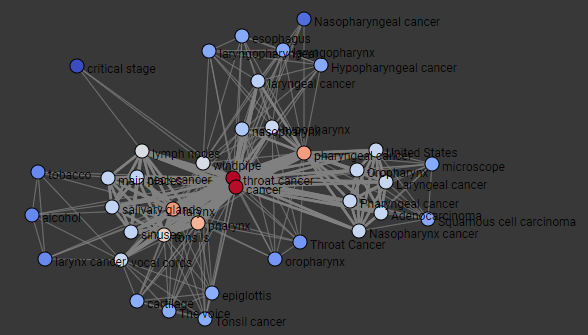
Baselines

Results analysis

Critical discussion

References

<https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sig-proceedings-template/bmvfhcdnxfty>



## 15/04

Info: Semana de Entrega do Projecto e Relatório

* Acabar ultimos pontos -> Fazer para mais do que 1 conversa/tópico.

Thresholds a calibrar que não estão na tabela:

* Na criação do grafo para remover entidades
* No stopping criteria do PR
* Peso das entidades das Queries vs das passagens

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Q entities penalty | Entities of the passages of the current turn  **DONE** | Entities of the passages returned in previous turns | Entities of the the passages of the first turn |
| 1 | 0 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 |
| 3 | 0 | 1 | 1 | 0 |
| 4 | 0 | 1 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 1 |
| 7 | 0 | 0 | 1 | 0 |
| 8 | 0 | 0 | 1 | 1 |
|  | X^2 |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

## 

## 22/04

Info: Determinar quais os objetivos para a próxima semana e iniciar o teste e implementação das novas variantes no peso dos grafos.

Objetivos:

Criar vetor de pergunta

Vai buscar entidades que tiverem 1 na primeira coluna;

Criar vetor de entidades de questão:

Iterar cada

## 29/04

Info: Iniciar e preparar o projecto

* Preparar o Colab
* Preparar os dados iniciais - conversas

Cenário:

* 20 conversas
* 8 perguntas em cada conversa
* 20 parágrafos candidatos a resposta possível a cada pergunta