Assignment 4: Search your transcripts. You will know it to be true. (Pt 2)

© Cristian Danescu-Niculescu-Mizil 2019

CS/INFO 4300 Language and Information

Due by 11:59pm on Wednesday February 19th

You must completely this assignment **individually.**

In this assignment we will build a basic information retrieval system using an inverted index and tf-idf representation. By the end of this assignment, you will have implemented all the necessary components to search through a collection of documents and return the most similar results to a given query.

Instructions

Follow the instructions below to get the necessary packages installed and set up your Python environment, then open the attached Jupyter notebook.

Run the notebook and complete the tasks contained in it, then upload the completed notebook and an HTML copy of it to CMS.

Double check that your files were correctly uploaded (by re-downloading them). If you have technical issues with CMS, send the files to the grad TAs and to the instructor via email **before** the assignment deadline, explaining what prevented you from submitting on CMS. As stated in the syllabus, we can not accept late submissions.

Learning Objectives

- Develop an understanding of the inverted index and its applications
- Explore use cases of boolean search
- Examine how the inverted index can be used to efficiently compute IDF values
- Introduce cosine similarity as an efficient search model

Academic Integrity and Collaboration

Note that these projects should be completed individually. As a result, all University-standard academic integrity guidelines must be followed.

\newpage

Setting up your environment

System Configuration

Perform the following steps in order:

1. Check your Version of Python3 (should be 3.7.6)

You can check via:

```
 > python3 --version
Python 3.7.6
```

If your version differs, then download `3.7` `here`.

2. Check that Pip is Installed and Up-to-date.

You should already have pip installed if you have Python downloaded from python.org. Make sure that yours is up-to-date.

Upgrade pip :

```
 > python3 -m pip install -U pip
```

If not, install it following instructions. (also found `here`)

```
 > curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py

> python3 get-pip.py
```

3. Download Virtualenv

`Virtualenv` helps establish an isolated `Python` environment. The environment allows you to separate project-specific dependencies and their versions from the `Python` modules installed locally on your computer. Once you have `virtualenv`, `cd` into the directory where the extracted assignment is stored (e.g. assignment1), and run:

```
 > virtualenv -p python3 venv
```

This creates a virtual environment called `venv`. In order to enter than virtual environment, run the following:

```
 > source venv/bin/activate
```

\newpage The following command line prompt will indicate that you're in the virtual environment:

```
 (venv) >
```

To deactivate the virtual environment, run the following:

```
 (venv) > deactivate
>
```

Whenever you work with this project, you should**always** be in your virtual environment. Without this isolation, we might run into module versioning issues and other problems when trying to run your project, which creates administrative overhead.

4. Install Dependencies

At the root of directory of the project skeleton code, run the following:

```
(venv) > pip3 install -r requirements.txt
```

This installs within your virtual environment all the necessary modules that are required at the beginning of the project.

5. Setup Jupyter Notebook

To use your virtualenv as the kernel for your Jupyter Notebook you run the following:

```
(venv) > python3 -m ipykernel install --user --name=venv
```

6. Open Jupyter Notebook and start working

Open the Jupyter Notebook enviroment and complete the assignment.

Make sure to go to Kernel > Change Kernel and click `venv` as the option.

```
 > source venv/bin/activate
(venv) > jupyter notebook
```