M2 NLP

UE 901 EC3: Intelligent Systems and Recommendations

Project Report - Part 1

Authors:
Asmaa Demny
Fatima Habib
Cécile Macaire
Chanoudom Prach
Ludivine Robert

January 25, 2021

Contents

1 Introduction

In the context of a job offer from Amazon, we want to build an efficient recommender system which consist mainly of "*filtering information that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications*". This project aims to build a recommender system that will take a Wikipedia page as an input, and will recommend 10 new links based on information that will be defined in this report.

The following report will present the main points:

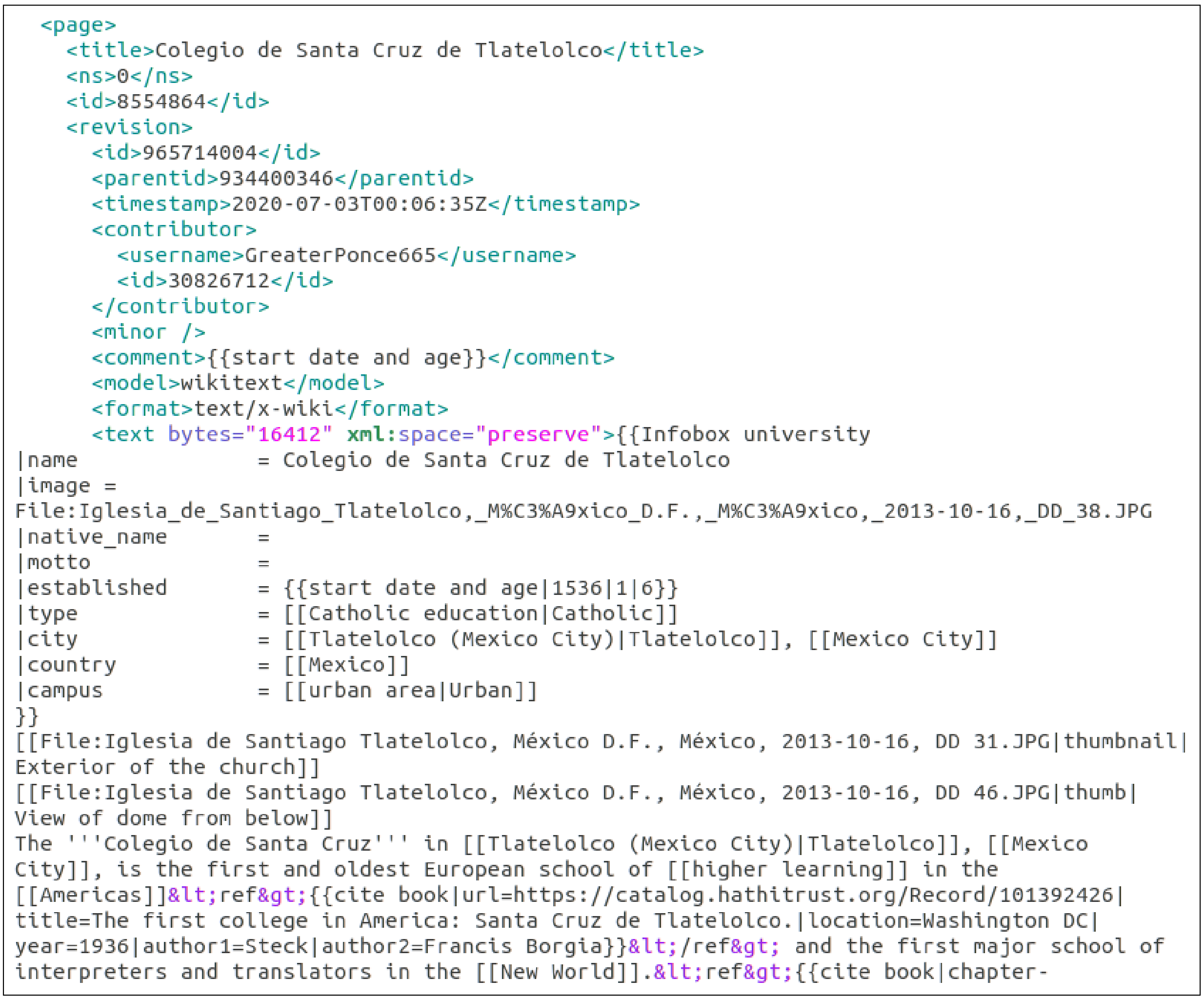
1. Data analysis and characterization,
2. Concurrence study (similar contexts),
3. Proposition of our recommender system strategy with the librairies,
4. The choosen approaches,
5. Analysis of the whole solution and conclusions.

2 Data analysis and characterization

From the ENWIKI DUMP PROGRESS ON 20210101[1] which is a Wikimedia dump service, we decided to take the data from:
`enwiki-20210101-pages-articles-multistream12.xml-p8554860p9172788.bz2`.
The following sections will provide information about the data structure, some statistics and the first ideas to answer the given problem.

2.1 Data structure & analysis

The data is a XML file of 166.3 MB. The XML file starts with a <mediawiki> tag which encompasses the metadata. The source language is given with the parameter *xml:lang*, here English. The recommender system will therefore only propose pages in this language.
The header of the file shows the site info in the tag <siteinfo> such as the site name ('Wikipedia'), the database name ('enwiki') and the namespaces. Each wiki page is included into a <page> tag.



```
  <page>
    <title>Colegio de Santa Cruz de Tlatelolco</title>
    <ns>0</ns>
    <id>8554864</id>
    <revision>
      <id>965714004</id>
      <parentid>934400346</parentid>
      <timestamp>2020-07-03T00:06:35Z</timestamp>
      <contributor>
        <username>GreaterPonce665</username>
        <id>30826712</id>
      </contributor>
      <minor />
      <comment>{{start date and age}}</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="16412" xml:space="preserve">{{Infobox university
|name              = Colegio de Santa Cruz de Tlatelolco
|image =
File:Iglesia_de_Santiago_Tlatelolco,_M%C3%A9xico_D.F.,_M%C3%A9xico,_2013-10-16,_DD_38.JPG
|native_name       =
|motto             =
|established       = {{start date and age|1536|1|6}}
|type              = [[Catholic education|Catholic]]
|city              = [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico City]]
|country           = [[Mexico]]
|campus            = [[urban area|Urban]]
}}
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 31.JPG|thumbnail|
Exterior of the church]]
[[File:Iglesia de Santiago Tlatelolco, México D.F., México, 2013-10-16, DD 46.JPG|thumb|
View of dome from below]]
The '''Colegio de Santa Cruz''' in [[Tlatelolco (Mexico City)|Tlatelolco]], [[Mexico
City]], is the first and oldest European school of [[higher learning]] in the
[[Americas]]&lt;ref&gt;{{cite book|url=https://catalog.hathitrust.org/Record/101392426|
title=The first college in America: Santa Cruz de Tlatelolco.|location=Washington DC|
year=1936|author1=Steck|author2=Francis Borgia}}&lt;/ref&gt; and the first major school of
interpreters and translators in the [[New World]].&lt;ref&gt;{{cite book|chapter-
```

Figure 1: Extract from the XML file of a Wikipedia page.

[1] `https://dumps.wikimedia.org/enwiki/20210101/`

From Figure 1, the most important information that we find for each Wikipedia page are:

– the title (<title>),

– the ID (<id>),

– the parent ID (<parentid>),

– the contributor (<contributor>),

– the text (<text>) which contains for some pages the infobox, the different sections identified by '==History==', and references (<ref>).

From the title page, we can access the corresponding Wikipedia page by adding underscores instead of spaces. For example, to access 'Colegio de Santa Cruz de Tlatelolco' page, the URL is `https://en.wikipedia.org/wiki/Colegio_de_Santa_Cruz_de_Tlatelolco`.

The XML file has 171742 Wikipedia pages. Only 5 pages have an empty content from the <text> tag. They will therefore not be used to build the recommender system.

2.2 Future work

Our first idea is to use Content-based filtering approach. Indeed, this approach uses content information about the items. It would consist of filtering the information inside the <text> tag of each Wikipedia page in order to generate links by the future recommender system. By information, we think of:

- infobox items,
- categories,
- related people,
- dates,
- places,
- movies,
- references cited,
- ...

We think about implemented our system using Python which has a lot of useful libraries for this purpose.