Pràctica de Programació Funcional + Orientada a Objectes
Similitud entre documents (Part II)

Programació Declarativa, Aplicacions.

2 de novembre de 2020

Introducció

Un cop sabem una manera de comparar documents de text, ens proposem aplicar-ho en l'entorn de la Vikipèdia. Analitzar la similitud entre pàgines web ens permetria, per exemple, detectar pàgines que s'assemblen molt però que no es referencien, o bé, pàgines que es referencien però no s'assemblen gens.

Per a dur a terme algun d'aquests anàlisis caldrà:

- **Poder comparar la similitud de dues pàgines:** per a calcular la similitud entre dues pàgines farem servir la del cosinus, però enlloc de considerar els vectors de *term frequencies (tf)* (freqüències de paraules), considerarem els vectors *tf_idf* que tenen en compte quant significativa és una paraula en el corpus (conjunt de pàgines) que estem analitzant. Fer servir nomes *tf* implicaria considerar que totes les paraules són igual d'importants. En el nostre cas, per exemple, si considerem pàgines del conjunt de pàgines que se us proporciona, la paraula "guerra" apareixerà a tots els documents ja que hem filtrat les viquis que contenen "segona guerra mundial".

  L'*inverse document frequency* (*idf*) d'un terme t en una col·lecció de documents C és: $idf = log\frac{|C|}{df}$ on $|C|$ és el nombre de documents de C i df és el nombre de documents on apareix el terme t. Fixeu-vos que quant en més documents apareixi, menys rellevant serà un terme per a fer comparacions.

  Ara ja podem definir *tf_idf* d'un terme t a un document d dins una col·lecció de documents C com a: $tf\_idf = tf \times idf$ on tf és el nombre de vegades que apareix t al document d i *idf* és l'inverse document frequency de t per la col·lecció de documents C.

  Compte, per a la similitud i per al càlcul dels *tf_idf* considerarem només les paraules que hi hagi a dins dels tags `<text>...</text>` eliminant les stop-words. **Cal eliminar més coses? links?**

- **Identificar les parts de les pàgines que ens són rellevants com ara el títol, el text i les referències que hi poden aparèixer.** L'aspecte d'un document XML de la Viquipèdia és el següent (en concret el primer de la nostra col·lecció, el `22.xml`):

```
<page>
    <title>Atletisme</title>
    <ns>0</ns>
    <id>22</id>
    <revision>
      <id>14061284</id>
      <parentid>13842377</parentid>
      <timestamp>2014-09-21T14:05:43Z</timestamp>
      <contributor>
        <username>Walden69</username>
        <id>1809</id>
      </contributor>
      <comment>/* Vegeu tambe */</comment>
      <text>{{millorar referencies|data=desembre de 2012}}
[[Fitxer:Athletics competitions.jpg|thumb|400px|Diferents competicions atletiques.]]

El mot atletisme prove de la paraula grega athlon que significa
lluita, competencia, combat o similars.
                        ...
{{ORDENA:Atletisme}}
[[Categoria:Atletisme| ]]
[[Categoria:Articles bons d'esport]]
[[Categoria:Articles bons dels 1000]]</text>
      <sha1>pmodgj07j924qtlj1xz7r7zcns5hg4e</sha1>
      <model>wikitext</model>
      <format>text/x-wiki</format>
    </revision>
  </page>
```

Ens convé saber identificar dues parts:

- El títol, que està entre els *tags* `<title>`*títol pàgina*`</title>`, i
- el contingut de la pàgina, que està entre els tags `<text>`*contingut*`</text>`.

A més, necessitarem saber com es referencien altres pàgines de la vikipèdia en general. El mecanisme és utilitzar el títol de la pàgina referenciada entre claus: `[[`*títol pàgina referenciada*`]]`. De totes maneres hi ha molts detalls a considerar:

- quan volem que es vegi un text però que es referenciï una pàgina en concret s'afegeix la barra `|` com a separadora, e.g. `[[Bombardeig de Guernica|Guernica]]`, enllaçarà amb la pàgina del bombardeig de Guernica però mostrarà la paraula Guernica com a enllaç.
- Per referenciar una secció de la mateixa pàgina, es fa servir `[[#`*secció*`]]`. També es pot referenciar una secció d'una altra pàgina: `[[`*títol pàgina*`#`*secció*`]]`.

- També hi ha enllaços a pàgines encara no definides: e.g. `[[MG 151 # MG 151/20|MG 151/20]]`. Evidentment no ens interessen.
- Quan es fa referència a un fitxer, com ara una imatge, es fa amb `[[Fitxer:`*nom_del_fitxer i opcions*`]]`. Per exemple `[[Fitxer:Bomber stream.jpg|thumb|300px|right|Part d'un transport ...]]`. Aquestes referències òbviament no les haurem de considerar ja que no referencien altres pàgines.

Hi ha molts més detalls que podeu consultar a: `http://en.wikipedia.org/wiki/Wiki_markup`

Què s'ha de fer?

En concret el que s'us demana és el següent:

1. Identificar les 100 (o més) pàgines més rellevants. Per fer això només tindrem en compte el nombre de referències que te cada pàgina, així, la pàgina amb més referències és la més important i la que en te menys és la menys important.
2. Detectar les pàgines que s'assemblen més però no es referencien mútuament. Podeu considerar que dues pàgines s'assemblen molt si la similitud tf_idf (considerant com a corpus les 100 pàgines més rellevants) és superior a 0.50.
3. Cal que useu `MapReduce` allà on hi veieu que hi pugui ajudar.
4. Cal que el `MapReduce` el parametritzeu amb un cert nombre de mappers i de reducers. Mostreu el temps que ha trigat el vostre procés sencer considerant 1, 4, 10 i 20 actors. És eficient provar amb més actors? fins quants?

Pistes

- A la modificació del `MapReduce` per tenir un cert nombre de mappers i reducers, us podeu inspirar en com es repartia el comptatge de primers entre els mappers i els reducers a l'exemple vist a classe.
- L'input dels `MapReduce` que han de treballar amb els fitxers hauria de ser una llista dels noms dels fitxers, de manera que les funcions de mapping ja fessin la feina de llegir i processar-los.
- La indicació d'agafar les **100 pàgines** més rellevants i el **0.50** de similitud són uns mínims. Quants més fitxers pogueu tractar millor. Podeu arribar a considerar les 5000 pàgines?

Lliurament

- El lliurament obligatori te com a data límit l'11 de desembre
- Les pràctiques es poden fer en equips de dos

- Cal lliurar un link GitHub amb el codi ben documentat
- Caldrà lliurar també un document on hi hagi:
  - extractes comentats del codi mostrant les principals funcions d'ordre superior usades (sobretot per la primera part), així com la part de la modificació del `MapReduce`
  - jocs de proves dels resultats obtinguts tant de la primera part com de la segona (llistat de les més importants amb nombre de referències i llistat de pàgines similars sense referències...)
  - com s'han fet tots els `MapReduce` (input, mapping i reducing)
  - quines classes us ha convingut crear i per què
  - taula de rendiment segons nombre d'actors en el `MapReduce`. Cal que especifiqueu la màquina feta servir
  - altres consideracions que volgueu ressaltar (com ara fins a quants documents heu pogut tractar)
- Es farà una presentació de la pràctica presencial/virtual al professor amb els dos membres de l'equip de la pràctica