Text Mining II - Diploma en Big Data (2014-2015)

Optimización PAN-AP-2013



Francisco Rangel Paolo Rosso



String+=



https://s3.amazonaws.com/cosmos_presentaciones/StringPlusEqual.mp4

StringBuilder.add



https://s3.amazonaws.com/cosmos_presentaciones/StringBuilder.mp4

String+= vs. StringBuilder

```
// Write doc frequencies: list of tf
StringBuilder Line = new StringBuilder();
sLine = sAuthor + "," + sLang + ",";
                                                                        245.333 términos
int length = oListOfTerms.size(); 
for (int i=0;i<length;i++) {</pre>
   String sTerm = oListOfTerms.get(i);
   if (oDoc.containsKey(sTerm)) {
                                                                        String+=
        int iFreq = oDoc.get(sTerm);
       sLine += iFreq;
    } else {
       sLine += "0";
     if (oDoc.containsKev(sTerm)) {
         int iFreq = oDoc.get(sTerm);
         Line.append(iFreq);
      } else {
         Line.append("0");
   if (i<oListOfTerms.size()-1) {</pre>
       sLine += ",";
   } else {
       sLine += "\n";
      if (i<length-1) {
                                                                        StringBuilder
         Line.append(",");
      } else {
         Line.append("\n");
 sLine += Line.toString();
oFW.write(sLine);
                                                                       ¿Es necesario?
oFW.flush();
```

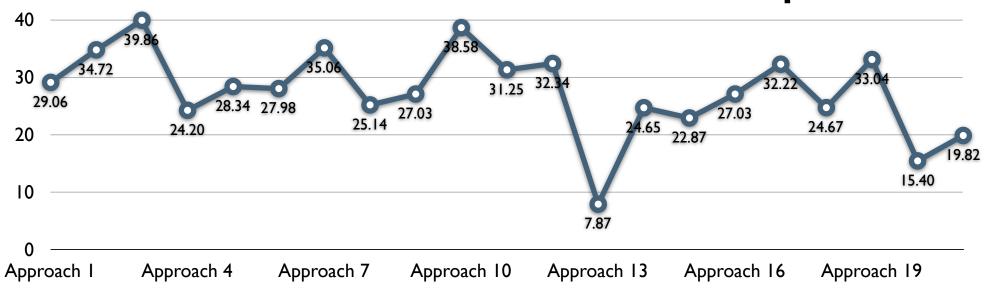
Big Data

```
3 UTC 2014 Omiguel_demetrio - RT @diana_markel379\nua yeocars
13 UTC 2014 @AutoExperto -
13 UTC 2014 @MicaelaMatto - RT @diana_marceta79#78A Yavanventent
13 UTC 2014 @khaleesicl_ 20RT @dianacharRelW7930Wnnjbo&K6B&CeaReA
13 UTC 2014 @una_intrusaymas QTRTdadtana@martel979Unudoyeochr&kée
   UTC 2014 @Genestldag - RT @diana_marcela79: Un joven británico
       2014 @DanielaAcostajiapaRTapdtana9haFde18790 bnitaoinobfell
        2014 @Casrolinamott - RT @diana_marcela79: Un joven britant
```

- ▶ Recuperar y almacenar
- **▶** Evolución
- ▶ Palabras y temas
- ▶ Etiquetado
- ▶ Hashtags
- **▶** Personas
- **▶** Localizaciones
- **▶** Marcas
- ▶ Tono, emoción
- ▶ Usuario, relaciones
- **▶** Influencia
- ▶Sexo y edad
- ▶ Perfil autor
- **)**...

https://s3.amazonaws.com/cosmos_presentaciones/ataque.mov

Precisión vs. coste computacional



IDENTIFICACIÓN DE EDAD Y GENERO EN ~240K AUTORES

