

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

Optimización PAN-AP-2013

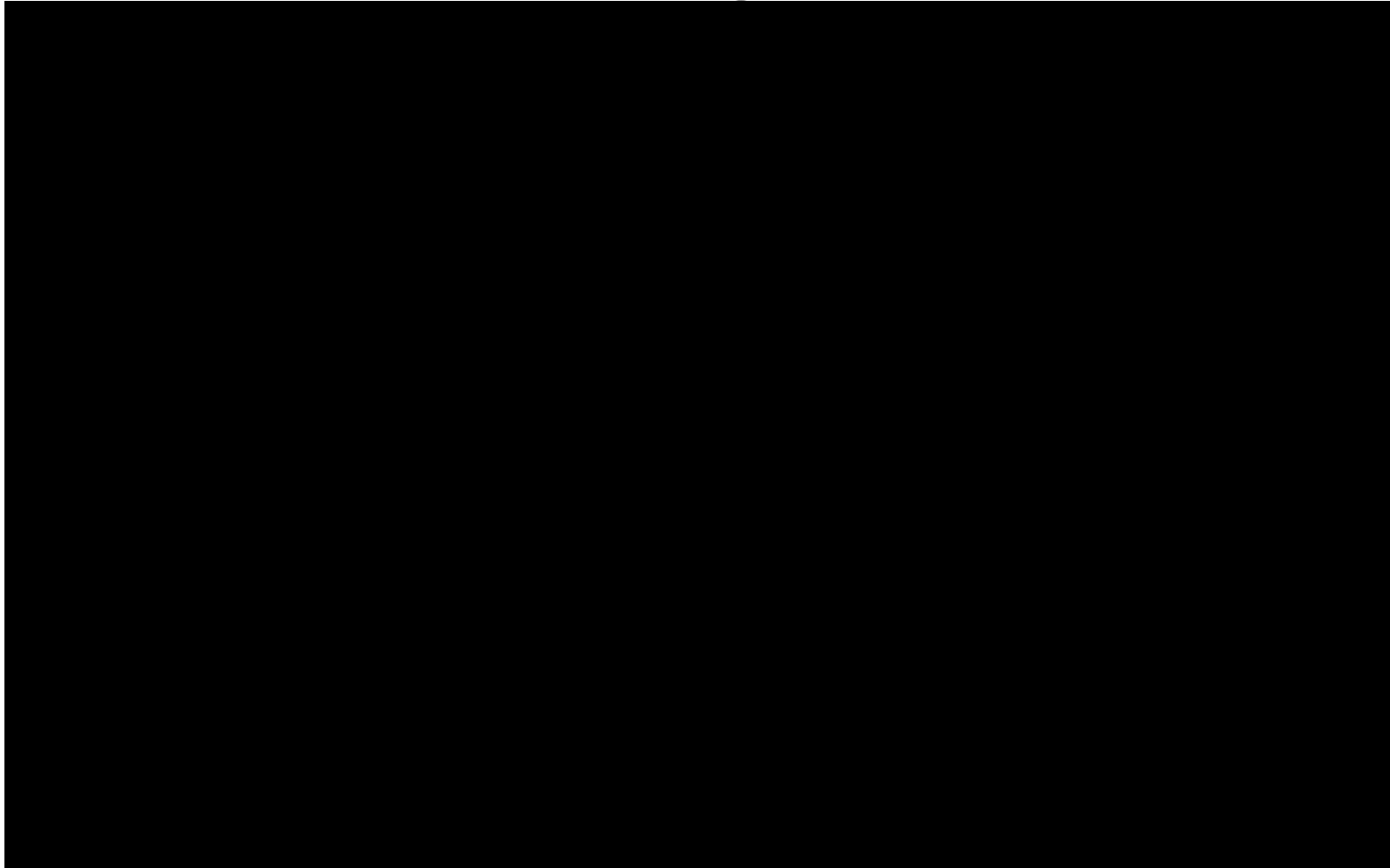


Francisco Rangel
Paolo Rosso



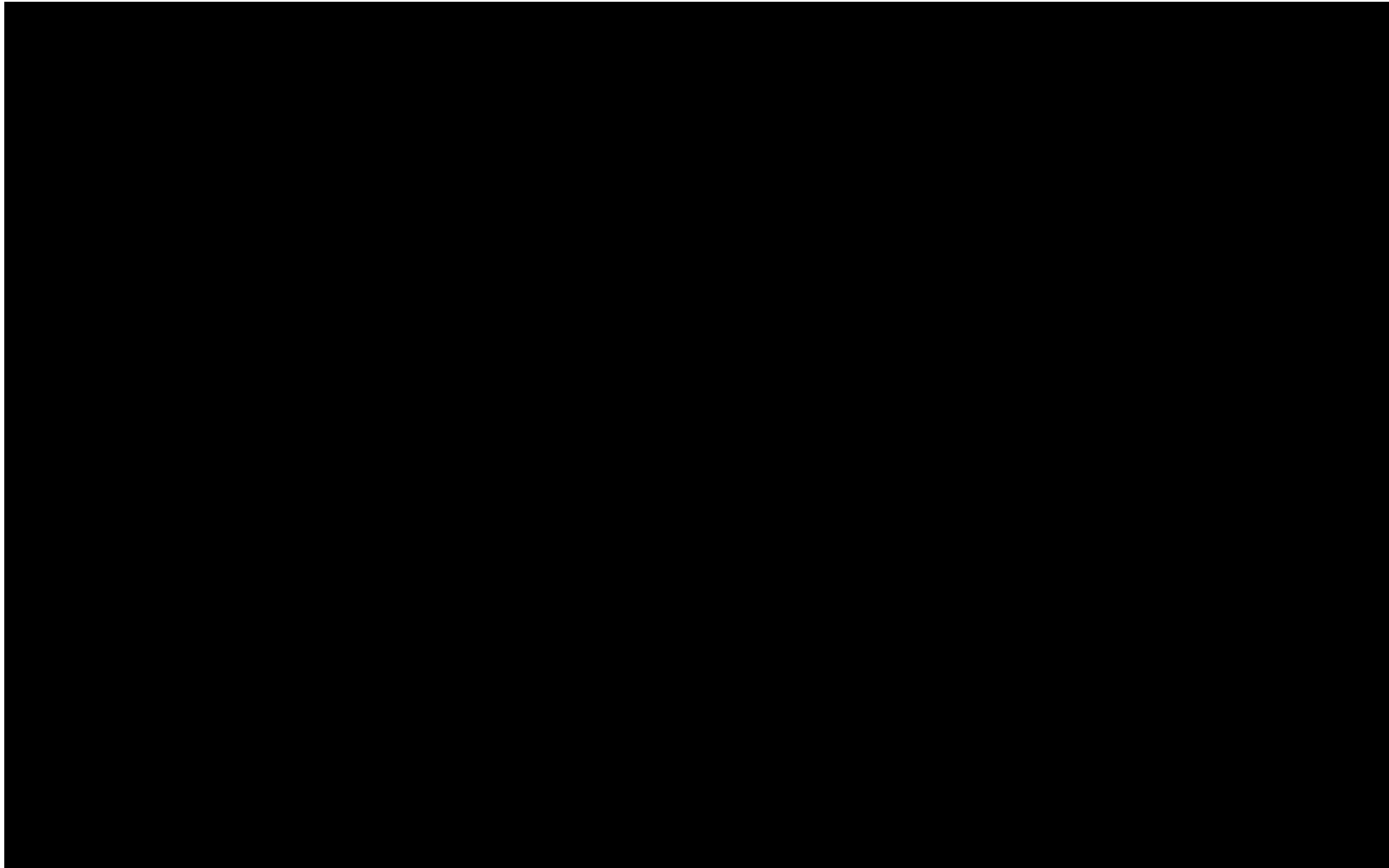
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

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 sLine = new StringBuilder();
sLine = sAuthor + "," + sLang + ",";
int length = oListOfTerms.size();
```

245.333 términos

```
String sTerm = oListOfTerms.get(i);
if (oDoc.containsKey(sTerm)) {
    int iFreq = oDoc.get(sTerm);
    sLine += iFreq;
} else {
    sLine += "0";
}
```

String+=

```
//          if (oDoc.containsKey(sTerm)) {
//              int iFreq = oDoc.get(sTerm),
//              Line.append(iFreq);
//          } else {
//              Line.append("0");
//          }
```

```
if (i<oListofTerms.size()-1) {
    sLine += ",";
} else {
    sLine += "\n";
}
```

```
//         if (i<length-1) {
//             Line.append(",");
//         } else {
//             Line.append("\n");
//         }
```

StringBuilder

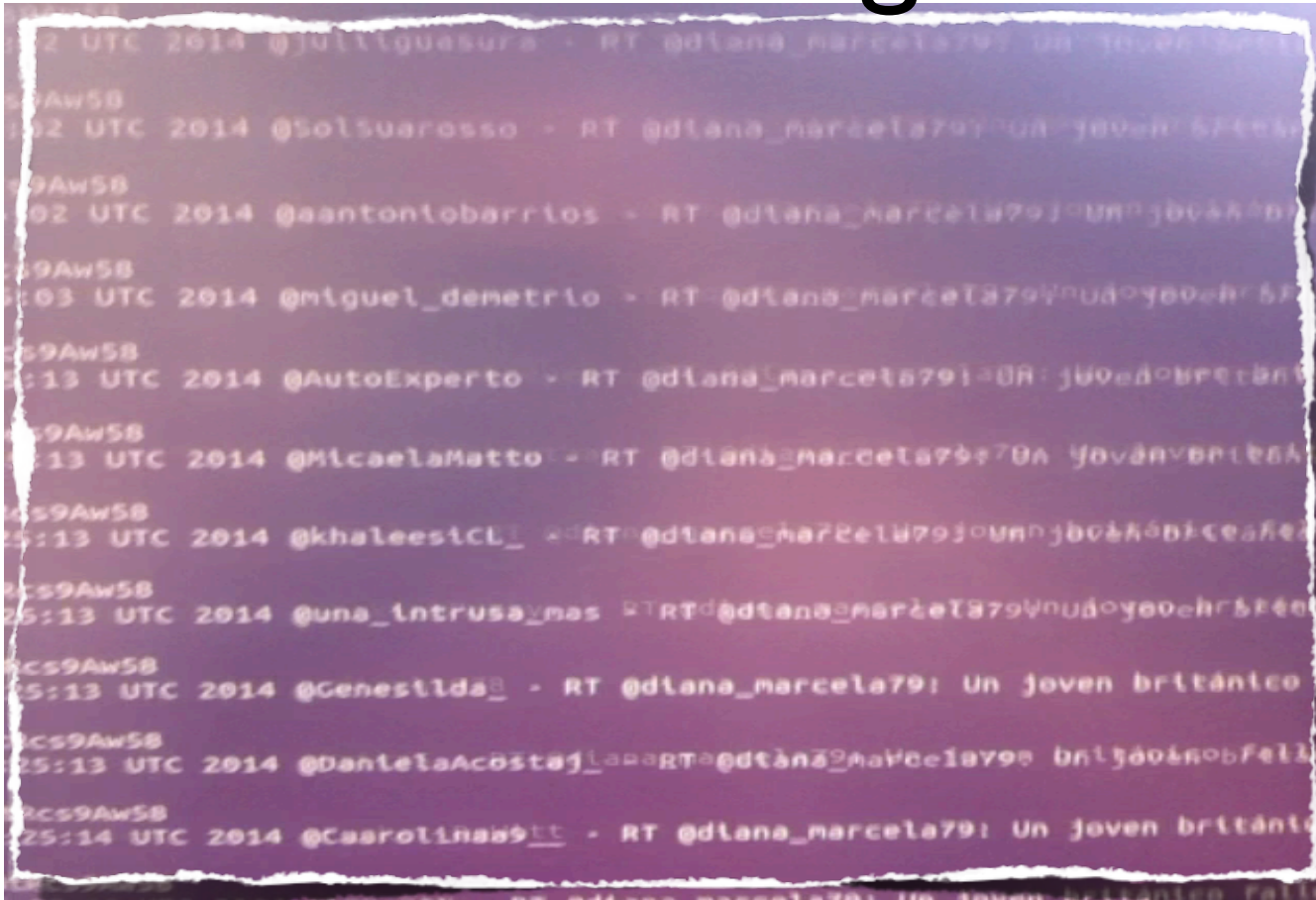
```

}
    sLine += Line.toString();
    oFW.write(sLine);
    oFW.flush();

```

¿Es necesario?

Big Data

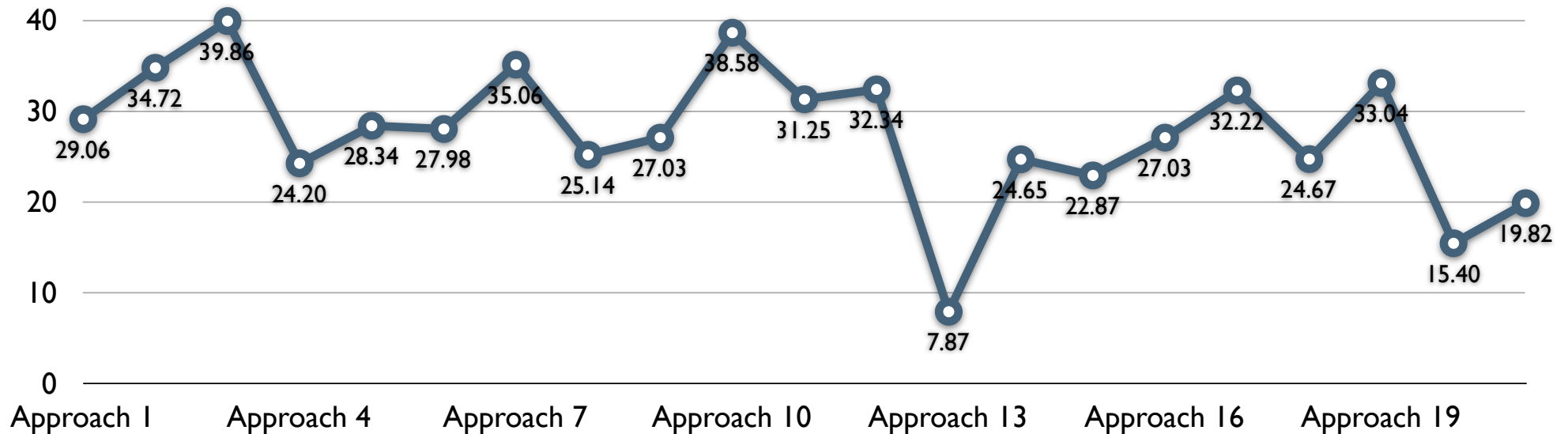


https://s3.amazonaws.com/cosmos_presentaciones/ataque.mov

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

| | | | |
|---------------|--------------|------------------|-----------------------|
| 80~120 | =4.800~7.200 | =288.000~432.000 | =6.912.000~10.368.000 |
| tuits/segundo | tuits/minuto | tuits/hora | tuits/día |

Precisión vs. coste computacional



IDENTIFICACIÓN DE EDAD Y GENERO EN ~240K AUTORES

