Eszközök az előfeldolgozáshoz

Apache Tika

- http://tika.apache.org/
- kinyeri a szöveget sokféle formátumú fájlból, pl. pdf, doc(x), xml, ppt...
- szabadon felhasználható, letölthető, nyílt forráskódú
- van hozzá python library: https://github.com/chrismattmann/tika-python

Bármi2UTF-8

- iconv
- python encode-decode
- stylo függvényei

Tokenizáló és mondatrabontó

- minden nyelvi elemző eszköz ezzel kezdődik --> lehetőségekért lásd a Nyelvi feldolgozó eszközök Google doksit

Nyelvdetektáló

- langdetect: az AVOBMAT-ba ezt építették be