Pipeline

Here is described the pipeline used for transforming the articles taken from wikipedia to nltk.Text objects to bigrams (using all the convenient methods from nltk library). For each entry e returned from wikipedia api the pipeline steps are:

1. entry = get_contents_for(e)
2. entry = [w['extract'] for w in entry]
3. entry = reduce(lambda l1, l2: l1 + l2, entry)
4. entry = entry.encode('ascii', 'ignore')
5. entry = entry.strip()
6. tokens = nltk.wordpunct_tokenize(entry)
7. bigrams = nltk.bigrams(tokens)
8. article = nltk.Text(tokens)