Analysing the Custom Tokenizer

Introduction

The custom tokenizer (CUS) was a tokenization model presented by D. Trieschnigg, W. Kraaij and F. de Jong for their paper on "The Influence of Basic Tokenization on Biomedical Document Retrieval". CUS combines compound tokenization, use of stop words and Porter Stemming. For this experiment the encoding for the Porter Stemmer utilized was developed by Brad Patton. For this project, CUS was modified to better adapt to the document collection and the search engine, thus ignoring special Greek characters such as Kappa (κ) or Gamma (γ).

How Does It Work?

Initially the text to be tokenized is transformed into a sequence of lowercase characters and split on whitespace. Each word is then stripped of any special character while numbers and letters are separated. As an example "international-no02" is split into the tokens "international", "no", "02". For each token we verify if they are contained within a list of 418 stop words, and, if they are the token is thrown away and will not be added to the inverted index nor the compounded token. Each token is then stemmed and then united once more to generate the compounded token. At the end of the process "International-No02" is divided into 4 tokens: "intern", "no", "02" and "internno02".

Example:

"International doctors on state-of-the-art hospital, placed 2nd place on 'Top Hospitals Magazine' receive on-the-job training to utilize new drug, B/CD28-responsive, used to treat cancer patients."

Compound Tokenization ➔ international, doctors, on, state, of, the, art, stateoftheart, hospital, placed,2, nd, 2nd, place, top, hospitals, magazine, receive, job, onthejob, training, to, utilize, new, drug, b, cd,28, responsive, bcd28responsive, used, treat, cancer, patients

Custom Tokenization ➔ intern, doctor, state, art, stateart, hospit, place, 2, nd, 2nd, top, magazin, receiv, job, train, util, new, drug, b, cd, 28, respons, bcd28respons, treat, cancer, patient

Results

Compared to the baseline run that utilizes the simple tokenizer we can see an improvement of 15.09%. While the improvement seems significant when we perform a paired t-test for the means on Excel, we obtain a "t-Stat" value of -1.1740, which within the critical range of +/- 2.0154, the "t-Critical two-tail" value. Due to the results obtained from the paired t-test it is not possible to reject the null hypothesis that the nDCG values are random samples from the same distribution.

	Baseline nDCG	CUS Tokenizer nDCG
Topic 401	0.2786	0.3760
Topic 402	0.3072	0.6127
Topic 403	0.7574	0.7523
Topic 404	0.1519	0.2022
Topic 405	0.1724	0.1628
Topic 406	0.6168	0.7570
Topic 407	0.5425	0.5060
Topic 408	0.4667	0.4558
Topic 409	0.2891	0.3010
Topic 410	1.0000	1.0000
Topic 411	0.3726	0.5720
Topic 412	0.7615	0.8300
Topic 413	0.1496	0.2560
Topic 414	0.3247	0.3439
Topic 415	0.3904	0.3904
Topic 417	0.7232	0.7569
Topic 418	0.3361	0.5946
Topic 419	0.6940	0.7877
Topic 420	0.8937	0.8622
Topic 421	0.2557	0.2883
Topic 422	0.6791	0.6834
Topic 424	0.2331	0.5830
Topic 425	0.6942	0.8566
Topic 426	0.1823	0.1836
Topic 427	0.3040	0.3806

	Baseline nDCG	CUS Tokenizer nDCG
Topic 428	0.3386	0.3281
Topic 429	0.5840	0.8838
Topic 430	0.6823	0.7499
Topic 431	0.3295	0.6735
Topic 432	0.1160	0.0817
Topic 433	0.1101	0.1061
Topic 434	0.4315	0.4252
Topic 435	0.3370	0.2711
Topic 436	0.3969	0.4060
Topic 438	0.4034	0.4629
Topic 439	0.1887	0.1896
Topic 440	0.8386	0.8338
Topic 441	0.6984	0.8497
Topic 442	0.2121	0.2217
Topic 443	0.4298	0.5770
Topic 445	0.4018	0.4373
Topic 446	0.2073	0.2340
Topic 448	0.2351	0.2155
Topic 449	0.0943	0.0883
Topic 450	0.6981	0.6941

Average Baseline nDCG:	**0.4291**
Average CUS Tokenizer nDCG:	**0.4939**

Figure 1 - Baseline VS. CUS Tokenizer Results

	Baseline nDCG	*CUS Tokenizer nDCG*
Mean	**0.4291**	**0.4939**
Variance	**0.0564**	**0.0664**
Observations	**45**	**45**
Pearson Correlation	**-0.1161**	
Hypothesized Mean Difference	**0**	
df	**44**	
t Stat	**-1.1740**	
P(T<=t) one-tail	**0.1234**	
t Critical one-tail	**1.6802**	
P(T<=t) two-tail	**0.2467**	
t Critical two-tail	**2.0154**	

Figure 2 - Paired t-Test results