Tokenizer いろいろ比較

- Solr 3.1で利用できるTokenizerを比較しました.

この資料のURL

- http://haruyama.github.com/Solr_5_20110516/
- http://goo.gl/nG5bG
  - ネヌジー5ビージー
- http://bit.l/ID4K1k
  - エルディ ー4ケー1ケー

自己紹介

- 春山 征吾 @haruyama
- 本日(2011/05/16) ECナビを退社.
- 転職先募集中!
  - 転職用情報 - 春山征吾のWiki

発表の流れ

- 評価手法
- Tokenizer(&Filter紹介)
- 比較
  - NGram系 vs 形態素解析
  - 3.1.0 vs 1.4.1

評価手法

- 日本のWikipediaのデータをupdate&commitした際の時間とインデックスサイズを比較
  - 元データは，TSV(4.7GB，1464241件)
- Solr/Tokenizer評価201105 - 春山征吾のWiki
- 1つの大きいTSVを突っ込んだ時の話
  - リソースが許せば，
    入力ファイルを分割して並列に更新することで
    時間を短縮できる

NGram系 vs 形態素解析

- NGram系
  - NGramTokenizer　は　1024文字しか処理しないのでCJKよりもサイズが小さくなっている.

	CJK	NGram(bi-gram)
時間(mm:ss)	13:45	9:05
サイズ(Gbyte)	7.37	6.75

- 形態素解析

	Japanese(ipadic)	Japanese(chasen)
時間(mm:ss)	36:53	51:45
サイズ(Gbyte)	6.75	7.12

3.1.0 vs 1.4.1

・3.1.0

	CJK	NGram(bi-gram)
時間(mm:ss)	13:45	9:05
サイズ(Gbyte)	7.37	6.75

・1.4.1

	CJK	NGram(bi-gram)
時間(mm:ss)	14:10	9:15
サイズ(Gbyte)	7.37	6.75

さいごに

- 転職先募集中!
  - 転職用情報 - 春山征吾のWiki
  - 懇親会やTwitter,
    メールなどでご連絡お願い致します.