Chi-Squared Test

This project provide a text feature selection method with chi-squared test.

The script could run in stand-alone mode or cluster mode by hadoop streaming.

---	cat	non-cat	sum over cats
with word	A▯	B▯	A+B
without word	C▯	D▯	C+D
sum	A+C▯	B+D▯	N

$$\chi^2 = \frac{N(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)}$$

$\chi^2 = \frac{(AD-BC)^2}{(A+B)(C+D)}$ (abbrev for in-cat scenario)

Input Format

cat'\t'segments

cat is class label in string while segments are space separeted words from a certain passage

eg:

sport'\t'well done MSN congrats to Barcelona

Output Format

cat'\t'word'\t'chi2'\t'A'\t'B'\t'C'\t'D'\t'st

st means positive or negative relative

Dict Format

file 'all_cat_segs_cnt' records the pre-computed number of passages of each cat with format:

cat'\t'count

eg:

fashion'\t'347882

sport'\t'2443297

Usage

stand-alone:

```
cat input_passage.tst | ./mapred_chi2.py m | sort | ./mapred_chi2.py r > output_chi2.tst
```

cluster:

Refer to run_chi2_uni.sh