Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle

Jasleen Kaur
Research Scholar[+],
Assistant Professor*,
[+]Uka Tarsadia University, Bardoli
*Shroff SR Rotary Institute of
Chemical Technology, Ankleshwar
Gujarat, India
+91-8128672321
sidhurukku@yahoo.com

Dr. Jatinderkumar R. Saini
Research Supervisior[+],
Professor, Director(I/C)[*],
[+]Uka Tarsadia University, Bardoli
*Narmada College of Computer
Application, Bharuch
Gujarat, India
+91-9687689708
saini_expert@yahoo.com

ABSTRACT
With advent of Unicode encoding, Punjabi language content, written using gurmukhi script as well as in shahmukhi script, is increasing day by day on internet. Processing textual information involves passing it to various pre-processing phases. Stop-word elimination is one such sub phase. 256 Gurmukhi stop words had been identified from poetry, stories and online material and passed to Punjabi stemmer. After stemming, 184 stemmed stop words were generated and these stemmed stop words were passed to transliteration phase. This led to generation of stop words in shahmukhi script. For the first time in scientific community dealing with computational linguistics and literature processing using NLP techniques, the list of 184 stop words of Punjabi language is released for public usage and further NLP applications. The presented list consists of stop words of Punjabi language with their Gurmukhi, Shahmukhi as well as Roman scripted forms.

Categories and Subject Descriptors
• **Computing methodologies~Natural language processing** • *Computing methodologies~Artificial intelligence* • Computing methodologies~Language resources

General Terms
Algorithms, Design, Human Factors, Languages.

Keywords
Gurmukhi, Natural Language Processing, Stop word, Shahmukhi, Punjabi.

1. INTRODUCTION
Pre-processing plays an important role in text mining area of computer science [8]. In order to prepare the data that can be used for mining useful information, data must be pre-processed. Pre-processing of text is done for mainly to extract useful features from text. Various pre-processing steps include noise removal, special symbol removal and stop word removal.

Stop words are words which does not have no significant semantic relation to the context in which they exist [8]. These are common words those that occur frequently in most of the documents in a given collection. They are extremely common words that do not provide any useful information to select documents. Thus, they must not be included as indexing terms. So these kinds of words must be eliminated from text because of two reasons. Firstly, it reduces the feature space of words and secondly, it increases the classifier accuracy.

Removal of stop words is needed in many natural language processing applications like classification, segmentation, spelling normalization and stemming. Eliminating such words from consideration early in automatic indexing speeds processing, saves huge amounts of space in indexes, and does not damage retrieval effectiveness.

2. LITERATURE SURVEY
Text classification is an active research area in information retrieval and natural language processing. A fundamental step in text classification is a list of 'stop' words (stop word list) that is used to identify frequent words that are unlikely to assist in classification and hence are deleted during pre-processing. Till now, many stop word lists have been developed for various foreign languages such as Chinese, Arabic, and English. This section provides details of various works done for identification of stop words in foreign languages.

Lili H. and Lizhu H. [16] had given a refined definition for stop words in Chinese text classification from a perspective of statistical correlation, then propose an automatic approach to extracting the stop word list in text classification based on the weighted Chi-squared statistic on 2*p contingency table. And then evaluated the list of stop words using accuracy obtained from text classification experiment. Yao and Zenwen [19] constructed a Chinese stop word list. Chinese English stop word list containing 1289 words was constructed by merging the classical stop word list with the stop words depending on the different domain of the text document corpus. Savoy [18] defined a general stop word list for those words which serve no purpose for retrieval, but are used very frequently in composing the documents. They establish a general stop word list for French. First, all the word forms appearing in their French corpora is sorted according to their frequency of occurrence and extract the 200 most frequently

WIR '16, March 21-22, 2016, Indore, India

DOI: http://dx.doi.org/10.1145/2909067.2909073

occurring words. Second, all numbers, plus all nouns and adjectives more or less directly related to with the main subjects of the underlying collections is removed. Third, some non information bearing words, even if they did not appear in the first 200 most frequent words are included. The suggested French general stop word list contains 215 words, and by using such a stop word list, the size of the inverted file was reduced by about 21% for one test collection, and about 35% for the second corpus. Myerson [16], used two statistical measures such as document frequency and chi-square for identification if stop words. Then, χ2(weighted Chi-squared statistic) was used to measure statistical correlation between a word and classification categories. X2for the words are calculated then ordered increasingly. Consecutively, the first word in the ordered list has the minimum value of weighted Chi-squared statistic, i.e. it has a higher document frequency and lesser correlations with all the categories. Chinese corpus of the Mayor's Public Access Line Project texts was used to evaluate and, compare results of classifiers.

Zheng and Gaowa [20] proposed a method for constructing stop-words list based on entropy calculation for Mongolian language. First, is to determine initial stop word lists then the entropy of every word is calculated and then ordered ascending to entropy. The second step is to combine results with the Mongolian part of speech to produce the final stop-word list. Zou et al. [21] used an aggregated model to measure both the word frequency characteristic by statistical model and its information characteristic by information model. The generated list was compared with other existing lists and showed an improvement over others. Elkhair [7] conducted a comparative study on the effect of stop words elimination on Arabic IR. Three stop lists were used in the comparison. General stop-list, corpus based stop-list and, a combined stop list. Alhadidi and Alwedyan [1] implemented a hybrid stop-word removal technique for Arabic language based on a dictionary and an algorithm. The proposed technique has been tested using a set of 242 Arabic abstracts chosen from the Proceedings of The Saudi Arabian National Computer conferences, and another set of data chosen from the Jordanian Alrai Newspaper.

Saini [17] has used a stop word list consisting of five categories of stop words namely, generic stop words, HTML stop words, noise stop words, domain stop words and miss-spelling stop words. The said stop words have been used by the researcher for processing of English un-structured documents scripted in Roman.

3. STOP WORDS FOR SHAHMUKHI SCRIPTED PUNJABI LANGUAGE DOCUMENTS

India is a multilingual country where a large number of languages are spoken in day to day life. But language families that dominate are Indo-Aryan (which is spoken in North Western Region) and Dravidian language family (spoken in southern region). Sino-Tibetan is one of minority language family (spoken in eastern region). Indo-Aryan Language Family mainly consists of Hindi, Gujarati, Bengali, Punjabi, Marathi, Urdu and Sanskrit languages. Dravidian Family, similarly, mainly consists of Telugu, Tamil and Kannada languages while the Sino-Tibetan Family consists mainly of Manipuri, Meithei and Himalayish Languages [14].

Punjabi is an *Indo-Aryan language* spoken by 102 million native speakers worldwide, making it the *10th most widely spoken language*. Punjabi is the most widely spoken language in Pakistan as a first language, the eleventh-most widely spoken in India, and the third-most spoken native language in the *Indian Subcontinent*. Punjabi is the fourth-most spoken language in the *United Kingdom* and third-most spoken native language in *Canada*. There are two ways to write Punjabi: Gurmukhi and Shahmukhi. The word Gurmukhi translates into "Guru's mouth",and Shahmukhi means "from the King's mouth". In the Punjab province of Pakistan, the script used is Shahmukhi and differs from the *Urdu alphabet* in having four additional letters. In the Indian state of Punjab, the Gurmukhī script is generally used for writing Punjabi [3].

In Punjabi language using gurmukhi script, 256 stop words were identified from poetries, stories and other online material [15]. Initially, 175 stop words are identified from various stories, news articles available online and 165 stops words are identified from poems collected in different categories discussed by Kaur J and Saini JR [15]. After the union of both the files, 256 unique stop words are identified from poems as well as news articles.

These identified stop words are stemmed to convert to its root form. Stemming is way of converting a written text into its root form [4].Gupta V. [9] developed different rules for handling stemming for verbs, adverbs and pronouns. For example in Punjabi language, word 'ਕੁੜੀਆਂ' [kudiya] 'girls' is converted to its root 'ਕੁੜੀ'[kudi] 'girl'. These stemming rules are manually applied to 256 stops words identified from poetry as well as other Punjabi documents. After applying these stemming rules to stop words obtained in the last step, 186 unique stop words are found.

On lieu of Punjabi Grammar and Part of Speech (POS) based word class categorization, these 186 stemmed stop words are categorized into 4 different word classes: Adverbs [6], Verbs [6], Pronouns [6], Conjunctions [2] and other miscellaneous words. Any word which is not suitable for first four categories is assigned to miscellaneous one. 99 different adverb forms, 40 different verbs, 26 pronouns, 7 conjunctions are identified from 186 stemmed stop words. And remaining 14 stop words are assigned to miscellaneous category [13].

All this work has been done in Punjabi language written using gurmukhi script. As explained earlier, in Pakistan, Punjabi language is also scripted using shahmukhi script. As there is unavailability of stop word list in Punjabi language written using shahmukhi script, these stemmed 184 gurumukhi stop words are transliterated to generate stop words in shahmukhi script. Transliteration is form of converting text present in one script into another script [5]. Gurmukhi to Shahmukhi transliteration system is designed by Punjabi university Patiala and is available online [11]. List of transliterated shahmukhi stop words is presented in Table I. This table consists of word in gurmukhi script followed by its transliterated form in shahmukhi script, which is followed by its transliteration and translation in Roman Script.

Table I. List of stop words in Gurmukhi, Shahmukhi and Roman script

S. No.	Word in Gurumukhi	Word in shahmukhi	Word in Roman	Meaning	S. No.	Word in Gurumukhi	Word in shahmukhi	Word in Roman	Meaning
1	ਇਸ	اِس	*[isa]*	this	2	ਜਿਸ	جس	*[jisa]*	who, what, which

3	ਵਿਚ	وچ	*[vica]*	in the	4	ਨ	ن	*[na]*	no
5	ਤਕ	تک	*[taka]*	up	6	ਹੁਣ	ہُن	*[huṇa]*	now
7	ਵੀ	وی	*[vī]*	too	8	ਜਿੰਨ?	جِنّاں	*[jināṁ]*	whom
9	ਉਤ?	اُتوں	*[othon]*	upon	10	ਨਾਲ	نال	*[nāla]*	with
11	ਨਹ?	نہ یں	*[nahīṁ]*	no	12	ਚਾਹੇ	چاہے	*[cāhē]*	either
13	ਭੀ	بھی	*[bhī]*	too	14	ਕਿਸ	کِس	*[kisa]*	what
15	ਵਲ?	ولوں	*[valōṁ]*	by	16	ਪਿਛ?	پِچھوں	*[pichōṁ]*	after
17	ਇਹ	ایہ	*[iha]*	this	18	ਏਧਰ	ایدھر	*[ēdhara]*	around
19	ਏ	اے	*[iha]*	this	20	ਨੂੰ	نوں	*[nū]*	to
21	ਜਦ?	جدوں	*[jadōṁ]*	when, while	22	ਅਜਿਹੇ	اجیہے	*[ajihē]*	such
23	ਕਈ	کئی	*[ka'ī]*	many	24	ਹੀ	ہی	*[hī]*	only
25	ਤੱਦ	تدّ	*[tada]*	then	26	ਕੇ	کے	*[kē]*	by
27	ਅੰਦਰ	اندر	*[andar]*	within	28	ਹ?	ہاں	*[hain]*	yes
29	? ਤੇ	اُتّے	*[utē]*	upon	30	ਬਹੁਤ	بہُت	*[bahuta]*	much
31	ਸਾਬੁਤ	سابُت	*[sābuta]*	complete	32	ਕਾਫ਼ੀ	کافی	*[kāfī]*	enough
33	ਕਦੀ	کدی	*[kadī]*	sometime	34	ਹੁਣੇ	ہُنّے	*[huṇē]*	now
35	ਨ?	نیں	*[nēṁ]*	the	36	ਲਈ	لئی	*[la'ī*	for
37	ਜੀ	جی	*[jī]*	respect	38	ਕਿ	کہ	*[ki]*	that
39	ਕਿਸੇ	کِسے	*[kisē]*	someone	40	ਮਗਰ	مگر	*[magara]*	behind
41	ਪੁਰਾ	پُورا	*[pūrā]*	complete	42	ਦਾ	دا	*[dā]]*	of
43	ਨੇ	نے	*[nē]*	the	44	ਤਰ?	طرحاں	*[tar'hāṁ]*	like
45	ਹੋਵੇ	ہووے	*[hovē]*	if	46	ਫੇਰ	پھیر	*[phēra]*	later
47	ਜੇਕਰ	جیکر	*[jēkar]*	just in case	48	ਵੇਲੇ	ویلے	*[vēlē]*	times
49	ਦੇ	دے	*[dē]*	of	50	? ਥੇ	اُتھے	*[othē]*	there
51	ਜਿਹੜਾ	جہڑا	*[jēhara]*	which	52	ਕਿਤੇ	کِتے	*[kitē]*	somewhere
53	ਬਾਅਦ	بعد	*[bā'ada]*	after	54	ਇੱਥੇ	اِتّھے	*[ithē]*	here
55	ਸਾਰਾ	سارا	*[sārā]*	all, whole	56	ਜਿਨ੍ਹੂੰ	جِنوں	*[jinhanu]*	whom
57	ਚ?	چوں	*[cho]*	out	58	ਜਦ	جد	*[jad]*	when
59	ਕਦੀ	کدی	*[kadē]*	never	60	ਵਾਙ	وانگ	*[vāṅga]*	like
61	ਸਭ	سب	*[sab]*	all	62	ਦੌਰਾਨ	دوران	*[doraan]*	during
63	ਤ?	تاں	*[tan]*	when	64	ਵਰਗਾ	ورگا	*[varagā]*	like
65	ਕਿ	کہ	*[ki]*	that	66	ਜੋ	جو	*[jō]*	that

67	ਲਾ	لا	*[la]*	to attach	68	ਕਰਕੇ	کرکے	*[karkē]*	because
69	ਪੂਰਾ	پُورا	*[pura]*	complete	70	ਬਿਲਕੁਲ	بالکُل	*[bilkul]*	absolutely
71	ਨਾਲੇ	نالے	*[naale]*	also	72	ਐਹੋ	ایہو	*[eho]*	such
73	ਤ?	توں	*[ton]*	from	74	ਕਉਣ	کَون	*[kaun]*	who
75	ਹੋਣਾ	ہونا	*[hona]*	be	76	ਫਿਰ	پھر	*[pher]*	then
77	ਪਾਸ?	پاسوں	*[paso]*	from	78	ਤਦ	تد	*[tad]*	then
79	ਜਿਹਾ	جیہا	*[jeha]*	little	80	ਕੋਲ?	کولوں	*[kolon]*	from
81	ਏਸ	ایس	*[ēs]*	this	82	ਕਿੰਨਾ	کِتّا	*[kina]*	how much
83	ਜਿਨ?	جِنہاں	*[jina]*	who	84	ਜਿਵ?	جِویں	*[jivē]*	such as
85	ਕੁਝ	کُجھ	*[kujh]*	some	86	ਹੇਠ?	ہیٹھاں	*[hethan]*	below
87	ਦੁਆਰਾ	دوارا	*[dobara]*	by	88	ਸਾਰੇ	سارے	*[sarē]*	all
89	ਸਦਾ	سدا	*[sada]*	forever	90	ਜਿੱਥੇ	جِتّھے	*[jithē]*	where
91	ਏਥੇ	ایتھے	*[ethē]*	here	92	ਕੋਈ	کوئی	*[koi]*	someone
93	ਬਾਰੇ	بارے	*[barē]*	about	94	ਕੀ	کی	*[ki]*	what
95	ਕਦ	کد	*[kad]*	when to	96	ਜੀ	جی	*[je]*	please
97	ਕਦੇ	کدے	*[kadē]*	never	98	ਦੀਆਂ	دِیاں	*[dī'āṁ*	of
99	ਹੋਏ	ہوئے	*[hoye]*	happen	100	ਚਲਾ	چلا	*[chala]*	goes
101	ਰਹੇ	رہے	*[rahē]*	are	102	ਲੈ	لَے	*[lai]*	take
103	ਬਣੈ	بنو	*[bano]*	become	104	ਆਖ	آکھ	*[aakh]*	say
105	ਦੇਣੀ	دینی	*[dēṇī]*	give	106	ਬਣ	بن	*[baṇa]*	made
107	ਪਿਆ	پیا	*[pi'ā]*	lying	108	ਕਰ	کر	*[kara]*	do
109	ਹੋਇਆ	ہویا	*[hō'i'ā]*	happened	110	ਪੈਣ	پَین	*[pain]*	falling
111	ਗਈ	گئی	*[ga'ī]*	gone	112	ਕਹਿ	کہِ	*[kēh]*	say
113	ਲਗ	لگ	*[laga]*	seem	114	ਚੁਕੇ	چُکے	*[chukē]*	-
115	ਹੁੰਦਾ	ہُندا	*[hudā]*	happen	116	ਕਿਹਾ	کیہا	*[keha]*	said
117	ਜਾਂਦਾ	جاندا	*[jāndā]*	going	118	ਕਰਵਾਈ	کروائی	*[karvayei]*	conducted
119	ਵੇਖ	ویکھ	*[vēkha]*	see	120	ਬਣਾਏ	بنائے	*[banaye]*	created
121	ਸੁਣ	سُن	*[suṇa]*	hear	122	ਕੀਤਾ	کِیتا	*[kitta]*	carried out
123	ਆਈ	آئی	*[ā'ī]*	occurred	124	ਜਾਵਣ	جاون	*[javan]*	going
125	ਸਕਦੇ	سکدے	*[sakdē]*	can	126	ਦੇਖ	دیکھ	*[dēkh]*	see
127	ਜਾਵੇ	جاوے	*[javē*	go	128	ਆਦਿ	آدی	*[ādi]*	so on
129	ਜਾਂਦਾ	جاندا	*[janda]*	going	130	ਲਿਆ	لیا	*[li'ā]*	taken

131	ਕਰਣ	کرن	*[karana]*	doing	132	ਆ	آ	*[ā]*	come
133	ਲਗਾਉਦ?	لگاؤداں	*[lagoda]*	not involving	134	ਰਿਹਾ	رہا	*[reha]*	going
135	ਆਵੇ	آوے	*[aavē]*	arrives	136	ਗਿਆ	گیا	*[geya]*	been
137	ਕਰੀ	کری	*[kari]*	do	138	ਉਠ	اُٹھ	*[otha]*	arise
139	ਲਾਇਆ	لایا	*[laeya]*	attach	140	ਰਹੀ	رہی	*[rahi]*	been
141	ਰਹਿ	رہ	*[reh]*	living	142	ਉਸਨੇ	اُسنے	*[usnē]*	he
143	ਉਹ	اوہ	*[uha]*	he, she	144	ਤੁਸ?	تُسیں	*[tusi]*	you
145	ਸ?	ساں	*[sāṁ]*	Was	146	ਮੇਰਾ	میرا	*[mera]*	my
147	ਸਭ	سبھ	*[sabha]*	All	148	ਉਸਦੀ	اُسدی	*[usdi]*	his
149	ਹਨ	ہَن	*[hana]*	Are	150	ਤੇਰਾ	تیرا	*[tera]*	your
151	ਤੂੰ	توں	*[tu]*	You	152	ਉਸ	اُس	*[us]*	his
153	ਸੀ	سی	*[si]*	Was	154	ਉਏ	اوئے	*[oyē]*	person
155	ਹੋ	ہو	*[ho]*	Are	156	ਆਪ	آپ	*[aap]*	you
157	ਤੈਨੂੰ	تینوں	*[tēnu]*	You	158	ਸਨ	سن	*[san]*	was
159	ਤੁਸ?	تُساں	*[tusa]*	You	160	ਮ?	میں	*[mein]*	i
161	ਹ?	ہیں	*[hain]*	are	162	ਤੁਸੀ	تُسی	*[tusi]*	you
163	ਹੈ	ہے	*[hai]*	is	164	ਅਸ?	اسیں	*[assi]*	we
165	ਆਪਣਾ	اپنا	*[apna]*	my	166	ਪਰ	پر	*[par[*	but
167	ਜੇ	جے	*[jē]*	if	168	ਤੇ	تے	*[tē]*	and
169	ਅਤੇ	اَتے	*[aatē]*	and	170	ਤ?	تاں	*[tāṁ]*	so
171	ਜ?	جاں	*[jāṁ]*	or	172	ਭਾਵ?	بھاویں	*[bhāvēm]*	*although*
173	ਕੁਲ	کُل	*[kal]*	total	174	ਅਗਲੀ	اگلی	*[aagali]*	next
175	ਵਗ਼ੈਰਾ	وغیرہ	*[vaġairā]*	etc	176	ਵਰਗ	ورگ	*[varg]*	category
177	ਰੱਖ	رکھ	*[rakh]*	put	178	ਆਮ	عامَ	*[āma]*	common
179	ਲੱਗ	لگ	*[laag]*	take	180	ਲਾ	لا	*[lā]*	apply
181	ਗੱਲ	گلّ	*[gal]*	thing	182	ਹਾਲ	حالَ	*[hāla]*	condition
183	ਪੀ	پی	*[pī]*	drink	184	ਇੱਕ	اِک	*[ek]*	one

4. CONCLUSION

Stop-words are functional and general words of the language that usually do not contribute to the semantics of the documents and have no read added value. The removal of such words contributes to the improvement of classifier efficiency. Punjabi language can be written using two different scripts, Gurumukhi and Shahmukhi. In this paper, 184 stemmed Gurumukhi stop words are presented in its transliterated (in Shahmukhi script) and translated (in Roman script) forms. The list presented here is released for public use for NLP in Shahmukhi scripted documents.

5. REFERENCES

[1] Alhadidi B. and Alwedyan, M. 2008. Hybrid Stop Word Removal Technique for Arabic language, *Egyptian Computer Science Journal.* 30,1.

[2] Article Overview of Punjabi Grammar accessed from http://punjabi.aglsoft.com/punjabi/learngrammar/?show=conjunction.

[3] Article Punjabi Language accessed from http://en.wikipedia.org/wiki/Punjabi_language. On November 2014.

[4] Article stemming accessed from http://en.wikipedia.org/wiki/Stemming.

[5] Article Transliteration accessed from *https://en.wikipedia.org/wiki/Transliteration*.

[6] Bhatia, Tej K. 1993. *Punjabi: A Cognitive-Descriptive Grammar.* Routledge Descriptive Grammar Series.

[7] El-Khair I. A., 2006 Effect of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study, *International Journal of Computing & Information Sciences,* 4, 3.

[8] Feldman R. and Sanger J. 2007 The text mining handbook, Cambridge university press.

[9] Gupta V. 2014 Automatic Stemming of Words for Punjabi Language. *Advances in Signal Processing and Intelligent Recognition systems, Advances in Intelligent Systems and Computing*, 264, 73-84.

[10] Gupta V. And Lehal G.S. 2011 Preprocessing Phase of Punjabi Language Text Summarization, *International Conference on Information System for Indian Languages*, 139, 250-253.

[11] Gurmukhi to Shahmukhi Transliteration System available at http://g2s.learnpunjabi.org/default.aspx.

[12] Hao, L. and Hao, L. 2008 Automatic Identification of StopWords in Chinese Text Classification, *International Conference on Computer Science and Software Engineering*, 2008.

[13] Kaur J., and Saini J.R 2015. POS based word class categorization of Gurumukhi language stemmed stop words. International Conference in Information Communication Technology for Intelligent System, *Smart Innovation in Smart Technology Springer*, November 2015(in print).

[14] Kaur J. and Saini JR, 2014 A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families. *International Journal of Data Mining and Emerging Technologies* ISSN 2249-3220, 4 , 2, 53-60.

[15] Kaur J., and Saini J.R., 2015 A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language. *International Journal of Data Mining and Emerging Technologies*; ISSN: 2249-3212 (eISSN: 2249-3220), 5, 2, 114-120.

[16] Myerson R.B., 1996 Fundamentals of social choice theory.

[17] Saini J.R. 2009 Self learning taxonomical classification system using vector space document analysis model for web text mining in UBE, Ph. D. Thesis under guidance of Desai A.A., accepted by the Department of Computer Science. VNSGU, Surat.

[18] Savoy J., 1999 A Stemming Procedure And Stopword List For General French Corpora, *Journal of the American Society for Information Science*, 50, 10, 944-952.

[19] Yao Z. and Ze-wen C. 2011, Research on the construction and filter method of stop-word list in text Preprocessing, *Fourth International Conference on Intelligent Computation Technology and Automation,* 2011.

[20] Zheng G. And Gaowa G., 2010 The Selection of Mongolian Stop Words, *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, 2010.

[21] Zou F. , Wang F. L., Deng X., Han S. and Wang L. S.,2006 Automatic Construction of Chinese Stop Word List. Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 1010-1015.