2018

CE306 - Information Retrieval Assignment 1

REGISTRATION NUMBER: 1501801

NAME: ADAM HEWITT

Contents

Specification

Usage:

```
python 1501801.py URL…
```

The file 1501801.py takes a URL as a single command-line argument. Returns a list of indexed strings from the web page's plaintext.

Functions:

getPage (URL) - Opens URL from string argument using the python Modules; Beautiful Soup and URLLib. Returns a BeautifulSoup 'soup' (a tree object containing the source code for the web page).

parsePage (soup) – Takes a BeautifulSoup soup as argument, strips the tags leaving web page's plaintext and adds in content from meta tags, stores it in a string and returns the string.

preProprecessing (parsedPageString) – Takes a string as argument, converts all characters to lower case and tokenises it using the word_tokenize function from NLTK on every word. Word_tokenize returns a list with each word as a separate list item.

posTag (t) – Takes list of strings and attaches a part of speech tag to each word using the built-in POS tagger from NLTK. Returns a list with the word and the tag in each element.

selection (taggedPage) – Takes a list of tagged strings and strips from it a list of stop words from NLTK.Corpus (See the Appendix: Stopwords). Then adds every string with the tags NN, VB, JJ to a new list. Then performs a frequency distribution analysis revealing the 15 most common words in the text. The 15 most common words from the frequency distribution are then removed from the tagged list.

stem (selectedWords) – Stem takes the now shortened, list of tagged words and performs the NLTK lemmatization function on it, returning a list of Lemmas which is the final output of indexed words for the URL given at the start in getPage. I chose the NLTK word_lemmatize function as it not only stems words but also morphologically changes them. For example, the word 'dogs' becomes 'dog' - an operation that could be achieved with a basic stemmer, word_lemmatize also groups words together based on their meanings i.e. words like 'is' and 'are' will be stemmed to 'be'.

callFuncs (URL) – Calls all the above functions procedurally starting with getPage and ending with stem. Takes the URL given from the command line and stores the source code into 'soup'. Then uses 'soup' as argument for parsePage, storing the result in 'pageText.' 'pageText' is then supplied as an argument to preProcessing, storing the result in tokenizedPage. This process goes on until we reach 'stems' which is the final list after it has gone through all of the stages mentioned above.

Once the program has finished running, the results of stem(selectedWords) is printed to the console and then the result for each stage is stored in a txt file with the naming scheme: "Output_WebPageTitle".

Output

URL: http://csee.essex.ac.uk/staff/udo/index.html

```
----------------------------------------------------------------------------------------------------

Indexed Words for URL: http://csee.essex.ac.uk/staff/udo/index.html

----------------------------------------------------------------------------------------------------



File 'Output_The Udo.txt' created...



['udo', 'udo', 'research', 'research', 'interest', 'member', 'language', 'computation', 'group', 'lac',
'essex', 'recent', 'lac', 'october', 'earlier', 'events:15th', 'lac', 'lac', 'lac', 'lac', 'lac', '10th', 'lac',
'research', 'interest', 'natural', 'language', 'processing', 'nlp', 'ir', 'implementation', 'technique', 'real',
'application', 'intelligent', 'directory', 'enquiry', 'assistant', 'ypa', 'example', 'going', 'year', 'extraction',
'structured', 'data', 'engineering', 'issue', 'played', 'major', 'role', 'making', 'ypa', 'usable', 'online',
'system', 'developing', 'technique', 'allow', 'extraction', 'conceptual', 'document', 'collection',
'utilization', 'knowledge', 'task', 'type', 'document', 'range', 'web', 'page', 'newspaper', 'article',
'form', 'structured', 'data', 'example', 'application', 'place', 'bt', 'mobile', 'workforce', 'look', 'log',
'analysis', 'study', 'current', 'project', 'ongoing', 'collaboration', 'five-year', 'esrc', 'large', 'grant',
'human', 'right', 'era', 'big', 'data', 'language', 'computation', 'group', 'fox/poesio/kruschwitz', 'play',
'significant', 'role', 'build', 'link', 'minority', 'right', 'group', 'watch', 'space', 'centre', 'doctoral',
'training', 'cdt', 'intelligent', 'game', 'game', 'intelligence', 'iggi', 'offering', 'phd', 'student',
'scholarship', 'year', 'funded', 'epsrc', 'supported', 'industrial', 'partner', 'first', 'student', 'admitted',
'october', 'october', 'fourth', 'cohort', 'started', 'september', 'year', 'find', 'student', 'profile', 'cdt',
'involves', 'university', 'york', 'university', 'essex', 'queen', 'mary', 'university', 'london', 'goldsmith',
'college', 'university', 'london', 'want', 'know', 'drop', 'email', 'three-year', 'knowledge', 'transfer',
'partnership', 'ktp', '2014-2017', 'involving', 'minority', 'right', 'group', 'essex', 'poesio/kruschwitz',
'aimed', 'new', 'system', 'support', 'civilian-led', 'monitoring', 'human-rights', 'violation', 'iraq',
'ceasefire', 'portal', 'major', 'outcome', 'corpus', 'violence', 'act', 'arabic', 'social', 'medium',
'academic', 'resource', 'described', 'detail', 'lrec', 'paper', 'successful', 'knowledge', 'transfer',
'partnership', 'ktp', 'starting', 'point', 'ongoing', 'collaboration', 'signal', 'medium', 'university', 'lead',
'pi', 'pleased', 'miguel', 'martinez-alvarez', 'joined', 'ktp', 'associate', 'head', 'research', 'signal',
'timeline', 'exciting', 'news', 'collaboration', 'miguel', 'named', 'business', 'leader', 'tomorrow', 'video',
'portrait', 'miguel', 'coverage', 'event', 'main', 'actor', 'april', 'best', 'demo', 'award', 'signal', 'demo',
'exciting', 'signal', 'announced', 'raised', 'growth', 'capital', 'top', 'tier', 'investor', 'see', 'techcrunch',
'signal', 'win', 'ktp', 'partnership', 'award', 'february', 'feature', "innovateuk's", 'home', 'page', 'new',
'ktp', 'confirmed', 'start', 'creating', 'annotated', 'resource', 'semantic', 'wikis', 'anawiki', 'epsrc', 'led',
'development', 'phrase', 'detective', 'annotation', 'game', 'aimed', 'creation', 'annotated', 'corpus',
'dali', 'build', 'finding', 'lesson', 'learned', 'phrasedetectives', 'recent', 'research/development',
'project', 'sensei', 'fp7', 'started', 'november', 'finished', 'one-line', 'aim', 'make', 'sense', 'human-
human', 'conversation', 'example', 'applying', 'advanced', 'natural-language', 'engineering',
'technique', 'anaphora', 'resolution', 'multi-document', 'summarization', 'read', 'use', 'sensei',
'technology', 'predict', 'brexit', 'result', 'prefer', 'read', 'article', 'polish', 'look', 'automatic',
'adaptation', 'knowledge', 'structure', 'assisted', 'seeking', 'autoadapt', 'involving', 'school',
'computer', 'science', 'electronic', 'engineering', 'robert', 'gordon', 'university', 'aberdeen', 'open',
'university', 'developed', 'evaluated', 'method', 'adapting', 'constructed', 'domain', 'model',
```

'population', 'user', 'browsing', 'behaviour', 'application', 'large-scale', 'evaluation', 'developed', 'method', 'seeking', 'scenario', 'interactive', 'browsing', 'focus', 'pebl-ai', 'follows', 'patient', 'experience', 'public', 'engagement', 'blog', 'pebl', 'pebl-ai', 'result', 'automated', 'interface', 'clinical', 'commissioner', 'group', 'ccgs', 'community', 'serve', 'development', 'career', 'path', 'framework', 'ktp', 'involving', 'school', 'computer', 'science', 'electronic', 'engineering', 'jobserve', 'world', 'first', 'internet', 'recruitment', 'service', 'development', 'intelligent', 'mail', 'server', 'ktp', 'involving', 'school', 'computer', 'science', 'electronic', 'engineering', 'active', 'web', 'aws', 'ipswich', 'based', 'company', 'devoted', 'cutting', 'edge', 'enterprise', 'system', 'learning', 'disability', 'data', 'infrastructure', 'esrc', 'markup-based', 'knowledge', 'extraction', 'intelligent', 'directory', 'enquiry', 'ypa', 'bt', 'current/recent', 'involvement', 'panel', 'chair', 'karen', 'spärck', 'jones', 'winner', 'announced', 'fernando', 'diaz', 'congratulation', 'original', 'call', 'member', 'read', 'irsg', 'autumn', 'irsg', 'newsletter', 'text', 'analytics', 'meetup', 'london', 'game', 'ai', 'meetup', 'london', 'industry', 'chair', 'sigir', 'full', 'paper', 'demo', 'games4nlp', 'lrec', 'co-chair', 'hcomp', 'ictir', 'sponsorship', 'co-chair', 'chiir', 'user', 'modeling/personalization/experience', 'track', 'wsdm', 'hlt', 'senior', 'tutorial', 'demo', "text2story'18", "newsir'18", 'lrec', 'scientific', 'swisstext', 'iswc', 'sigir', 'full', 'paper', 'demo', 'tutorial', 'outstanding', 'reviewer', 'award', 'ranlp', 'fdia', 'paper', 'mindtrek', 'data-driven', 'gamification', 'design', 'worksop', 'cikm', 'iswc', 'games4nlp', 'senior', 'paper', 'demo', 'industry', 'co-chair', 'www', 'user', 'modeling/personalization/experience', 'track', 'eswc', 'co-chair', 'nlp+ir', 'track', 'multiling', 'chiir', 'mmm', 'gamification', "gamifir'16", 'sigir', 'sigir', 'full', 'paper', 'demo', 'air', 'paper', 'cikm', 'accessing', 'cultural', 'heritage', 'scale', 'jcdl', 'recent', 'advance', 'news', "newsir'16", 'konvens', 'reviewing', 'board', 'wsdm', 'chiir', 'senior', 'pc', 'demo', 'lrec', 'scientific', 'achs', 'jcdl', 'multiling', 'cikm', 'ir', 'track', 'air', 'jcdl', 'tutorial', 'co-chair', 'sigir', 'full', 'paper', 'fdia', 'gamification', "gamifir'15", 'senior', 'panellist', 'student', 'program', 'clef', 'news', 'recommender', 'challenge', 'newsreel', 'steering', 'eswc', 'hybrid', 'approach', 'translation', 'hytra', 'ranlp', 'ieee/wic/acm', 'international', 'web', 'intelligence', 'ir', 'sigir', 'sigir', 'symposium', 'ir', 'practice', 'sirip', 'gamification', "gamifir'14", 'single-shot', 'text', 'query', 'bridging', 'gap', 'research', 'community', "mindthegap'14", 'iconference', 'iiix', 'irf', 'lrec', 'scientific', 'eacl', 'hybrid', 'approach', 'translation', 'hytra', 'umap', 'personalised', 'multilingual', 'access', 'clef', 'news', 'recommender', 'challenge', 'newsreel', 'steering', 'konvens', 'reviewing', 'board', 'mmm', 'demo', 'co-chair', 'aaai', 'student', 'program', 'ieee/wic/acm', 'international', 'web', 'intelligence', 'sigir', 'paper', 'demo', 'co-chair', 'sigir', 'bar', 'sigir', 'enrich', 'ranlp', 'ijcnlp', 'dir', 'irf', 'ieee/wic/acm', 'international', 'web', 'intelligence', 'ir', 'student', 'program', 'hybrid', 'approach', 'translation', 'hytra', 'future', 'direction', 'access', 'fdia', 'sigir', 'industry', 'track', 'co-chair', 'industry', 'track', 'presentation', 'online', 'iiix', 'industry', 'query', 'session', 'sir2012', 'irf', 'konvens', 'reviewing', 'board', 'student', 'research', 'doctoral', 'consortium', 'eacl', 'joint', 'hybrid', 'machine', 'translation', 'lrec', 'creating', 'cross-language', 'resource', 'disconnected', 'language', 'style', 'ieee/wic/acm', 'international', 'web', 'intelligence', 'stair', 'industry', 'co-chair', 'future', 'direction', 'access', 'fdia', 'content', 'analysis', 'track', 'irf', 'invited', 'panelist', 'acl', 'hlt', 'student', 'session', 'hypertext', 'personalised', 'multilingual', 'hypertext', 'pmhr', 'ieee/wic/acm', 'international', 'web', 'intelligence', 'missed', 'read', 'article', 'itnow', 'acm', 'cikm', 'demo', 'co-chair', 'co-chair', 'missed', 'report', 'sigir', 'forum', 'industry', 'iiix', 'stair', 'irf', 'student', 'research', 'ranlp', 'minucs', 'ictir', 'future', 'direction', 'access', 'fdia', 'special', 'issue', 'journal', 'natural', 'language', 'engineering', 'interactive', 'qa', 'guest', 'editorial', 'board', 'corpus', 'profiling', 'co-located', 'interaction', 'context', 'iiix', 'organising', 'high-level', 'extraction', 'future', 'direction', 'access', 'fdia', 'flatlands', 'organising', 'organised', 'essex', 'language', 'computation', 'group', 'sigir', 'demonstration', 'ranlp', 'teaching', 'ce306/ce706', 'text', 'analytics', 'ce807', 'phd', 'student', 'co-supervised', 'deirdre', 'lungley', 'mahmoud', 'el-haj', 'm-dyaa', 'albakour', 'suma', 'adindla', 'sharhida', 'saad', 'azhar', 'alhindi', 'jon', 'chamberlain', 'roseline', 'antai', 'fawaz', 'alarfaj', 'maha', 'althobaiti', 'silviu', 'paun', 'an', 'alghamdi', 'david', 'gundry', 'external', 'co-supervisor', 'rob', 'homewood', 'external', 'co-supervisor', 'chris',

```
'madge', 'dino', 'ratcliffe', 'steve', 'zimmerman', 'publication', 'selected', 'publication', 'foundation',
'trend', 'book', 'searching', 'enterprise', 'review', 'martin', 'white', 'intranetfocus', 'springer', 'book',
'intelligent', 'document', 'review', 'informer', 'newsletter', 'bcs', 'irsg', 'review', 'acm', 'computing',
'review', 'udo', 'kruschwitz', 'university', 'essex', 'school', 'computer', 'science', 'electronic',
'engineering', 'wivenhoe', 'park', 'colchester', 'co4', 'tel', '+44-1206-872669', 'fax', '+44-1206-872788',
'email', 'udo', 'essex.ac.uk', '1997-2018', 'udo', 'last', 'change', 'january', 'udo', 'web', 'technology',
'engine', 'ir', 'knowledge', 'extraction', 'ontology', 'research', 'intelligent', 'document', 'springer',
'book', 'natural', 'language', 'engineering', 'udo']
```

URL: http://irsg.bcs.org/ksjaward.php

```
----------------------------------------------------------------------------------------------------

Indexed Words for URL: http://irsg.bcs.org/ksjaward.php

----------------------------------------------------------------------------------------------------


File 'Output_KSJ Award.txt' created...


['ksj', 'specialist', 'group', 'home', 'committee', 'join', 'contact', 'membership', 'informer', 'ecir',
'industry', 'award', 'solution', 'agm', 'committee', 'meeting', 'group', 'resource', 'tweet', 'bcs_irsg',
'irsg', 'supported', 'event', 'past', 'event', 'informer', 'read', 'irsg', 'activity', 'latest', 'edition',
'informer', 'apply', 'intrested', 'becoming', 'professional', 'join', 'bcs-irsg', 'web', 'badge', 'member',
'irsg', 'show', 'affiliation', 'displaying', 'web', 'badge', 'page', 'click', 'code', 'bcs/bcs', 'irsg', 'spärck',
'commemorate', 'spärck', 'british', 'computer', 'society', 'specialist', 'group', 'irsg', 'conjunction',
'created', 'commemorate', 'achievement', 'spärck', 'professor', 'emerita', 'computer', 'cambridge',
'remarkable', 'woman', 'computer', 'science', 'field', 'ir', 'natural', 'language', 'processing', 'nlp',
'regard', 'experimentation', 'outstanding', 'influential', "karen's", 'achievement', 'resulted',
'receiving', 'number', 'prestigious', 'accolade', 'lovelace', 'medal', 'advancement', 'system', 'acm',
'salton', 'significant', 'sustained', 'continuing', 'irsg', 'grateful', 'sponsorship', 'spärck-jones',
'sponsered', 'spärck', 'pleased', 'announce', 'bcs/bcs', 'irsg', 'spärck', 'awarded', 'fernando', 'diaz',
'director', 'independent', 'expert', 'ir', 'natural', 'language', 'processing', 'nlp', 'impressed', 'great',
'breadth', 'excellence', 'raising', 'important', 'issue', 'ethical', 'issue', 'service', 'community', 'addition',
'fact', 'fernando', 'made', 'major', 'ir', 'nlp', 'spirit', 'congratulation', 'looking', 'inspiring', 'talk', 'ecir',
'grenoble', 'spärck', 'difficult', 'task', 'decide', 'range', 'candidate', 'outstanding', 'credential', 'lead',
'impressive', 'programme', 'view', 'advance', 'understanding', 'ir', 'and/or', 'experimentation', 'giving',
'jaime', 'teevan', 'affiliate', 'associate', 'professor', 'washington', 'seattle', 'usa', 'principal',
'researcher', 'redmond', 'recognises', 'jaime', 'strong', 'creative', 'intersection', 'user', 'experience',
'social', 'medium', 'particular', 'jaime', 'advanced', 'understanding', 'design', 'implementation',
'careful', 'evaluation', 'new', 'algorithm', 'user', 'experience', 'jaime', 'give', 'recipient', 'spärck',
'impressed', 'quality', 'diversity', 'field', 'candidate', 'faced', 'hard', 'time', 'narrowing', 'nomination',
'short', 'list', 'turned', 'experienced', 'genuine', 'difficulty', 'deciding', 'candidate', 'best', 'matched',
'criterion', 'endeavoured', 'advance', 'understanding', 'ir', 'and/or', 'experimentation', 'consequence',
'unique', 'decided', 'make', 'full', 'award', 'alphabetical', 'order', 'jordan', 'boyd-graber', 'assistant',
'professor', 'computer', 'science', 'department', 'colorado', 'recognised', 'jordan', 'strong', 'creative',
'probabilistic', 'topic', 'model', 'particular', 'shown', 'semantic', 'coherence', 'topic', 'assessed',
'linguistic', 'knowledge', 'word', 'sens', 'syntactic', 'relation', 'inter-language', 'relationship',
'incorporated', 'user', 'feedback', 'integrated', 'learned', 'topic', 'model', 'improve', 'semantic',
'coherence', 'topic', 'thought', 'jordan', 'shaped', 'understanding', 'nlp', 'new', 'important', 'way',
'experimental', 'machine', 'learning', 'jordan', 'gave', 'ecir', 'interactive', 'machine', 'learning',
'understanding', 'large', 'document', 'collection', 'emine', 'yilmaz', 'lecturer', 'department',
'computer', 'science', 'college', 'london', 'impressed', 'body', 'emine', 'evaluation', 'technique',
'match', 'reality', 'modern', 'system', 'particular', 'missing', 'user', 'judgment', 'proposed', 'new',
'metric', 'robust', 'missing', 'data', 'expected', 'browsing', 'utility', 'new', 'evaluation', 'metric',
'derived', 'interaction', 'pattern', 'real', 'web', 'session', 'recent', 'focused', 'devising', 'evaluating',
'quality', 'task', 'based', 'system', 'system', 'help', 'user', 'complete', 'task', 'led', 'issue', 'query',
'engine', 'opposed', 'retrieving', 'list', 'document', 'relevant', 'query', 'submitted', 'recognized',
'emine', 'pushed', 'field', 'ir', 'forward', 'important', 'aspect', 'high', 'impact', 'academia', 'practice',
'emine', 'gave', 'ecir', 'task-based', 'perspective', 'spärck', 'ryen', 'white', 'senior', 'researcher',
'redmond', 'recognises', 'ryen', 'white', 'contributed', 'many', 'layer', 'better', 'understanding',
'interaction', 'eg', 'novel', 'creative', 'analysis', 'vital', 'unexplored', 'aspect', 'searcher', 'behavior',
'coupled', 'development', 'model', 'application', 'improving', 'experience', 'impressed', 'ryen',
'academic', 'output', 'many', 'leading', 'sole', 'author', 'using', 'experimental', 'method', 'large-scale',
'log', 'analysis', 'user', 'study', 'survey', 'sheer', 'productivity', 'take-up', 'academia', 'number', 'best',
'paper', 'award', 'prestigious', 'conference', 'including', 'acm', 'sigir', 'acm', 'cikm', 'acm', 'sigchi',
'jasist', 'speak', 'hallmark', 'ryen', 'professional', 'activity', 'continued', 'serve', 'academic',
'community', 'reviewing', 'track', 'chairing', 'pc', 'co-chairing', 'top', 'conference', 'editorial', 'board',
'membership', 'guest', 'editing', 'researcher', 'industry', 'ryen', 'gave', 'ecir', 'mining', 'modelling',
'online', 'health', 'slide', 'lecture', 'available', 'eugene', 'agichtein', 'associate', 'professor',
'mathematics', 'computer', 'science', 'department', 'emory', 'atlanta', 'usa', 'making', 'bcs/', 'irsg',
'spärck-jones', 'recognizes', 'eugene', 'agichtein', 'made', 'several', 'important', 'large', 'scale', 'web',
'text', 'data', 'mining', 'focus', 'user', 'interaction', 'data', 'eugene', 'included', 'demonstrating',
'model', 'human', 'interaction', 'inferred', 'leveraging', 'computational', 'technique', 'large-scale',
'behaviour', 'record', 'influential', 'depth', 'several', 'area', 'impressive', 'eugene', 'focused',
'understanding', 'modelling', 'user', 'interaction', 'web', 'collaborative', 'question', 'answering',
'example', 'shown', 'click', 'form', 'implicit', 'feedback', 'useful', 'improving', 'result', 'ranking',
'gathered', 'large', 'number', 'user', 'area', 'contribution', 'seen', 'influential', 'eugene', 'includes',
'movement', 'prediction', 'response', 'advertising', 'reached', 'domain', 'overall', 'eugene', 'agichtein',
'regarded', 'excellent', 'experimentalist', 'recognizes', 'critical', 'linkage', 'theory', 'experiment',
'eugene', 'gave', 'speech', 'diane', 'associate', 'professor', 'north', 'carolina', 'nc', 'usa', 'making', 'bcs/',
'irsg', 'spärck-jones', 'recognizes', 'diane', 'made', 'important', 'analysis', 'seeking', 'behavior',
'development', 'new', 'experimental', 'method', 'system', 'support', 'seeking', 'analysis', 'diane',
'made', 'several', 'important', 'user', 'modeling', 'using', 'implicit', 'indicator', 'relevance',
'development', 'analysis', 'interface', 'elicit', 'richer', 'statement', 'interest', 'new', 'methodology',
'designing', 'evaluating', 'interactive', 'system', 'strong', 'user-oriented', 'view', 'users-as-people',
'cognitive', 'task', 'diane', 'gave', 'speech', 'slide', 'talk', 'available', 'download', 'made', 'evgeniy',
'gabrilovich', 'senior', 'scientist', 'manager', 'nlp', 'ir', 'group', 'yahoo', 'california', 'u.s.', 'evgeniy',
'gave', 'speech', 'abstract', 'talk', 'bio', 'available', 'first', 'ksj', 'mirella', 'lapata', 'reader', 'assoc', 'prof',
'school', 'informatics', 'edinburgh', 'mirella', 'focused', 'various', 'problem', 'emphasis', 'statistical',
'method', 'text', 'generation', 'application', 'worked', 'complex', 'problem', 'word', 'sense',
'disambiguation', 'ambiguity', 'resolution', 'semantic', 'vector', 'space', 'story', 'generation', 'many',
'others', 'abstract', 'talk', 'ecir', 'bio', 'available', 'download', 'previous', 'member', 'list', 'previous',
'member', 'alphabetical', 'order', 'nicholas', 'belkin', 'rutgers', 'pia', 'copenhagen', 'ann', 'copestake',
'cambridge', 'susan', 'dumais', 'rob', 'gaizauskas', 'sheffield', 'ayse', 'goker', 'robert', 'gordon', 'chair',
'2009-2013', 'katja', 'hofmann', 'cambridge', 'joemon', 'jose', 'glasgow', 'udo', 'kruschwitz', 'essex',
'chair', 'rada', 'mihalcea', 'michigan', 'marie-francine', 'moens', 'katholieke', 'universiteit', 'doug',
'oard', 'maryland', 'carol', 'peter', 'cnr', 'pisa', 'stephen', 'robertson', 'city', 'college', 'london', 'stefan',
'rueger', 'open', 'uk', 'chair', '2014-2016', 'tomek', 'strzalkowski', 'suny', 'albany', 'bonnie', 'webber',
'edinburgh', 'marti', 'hearst', 'california', 'berkeley', 'mark', 'sanderson', 'rmit', 'milad', 'shokouhi',
'copyright', 'legal', 'privacy', 'notice', 'registered', 'charity', 'page', 'last', 'modified', 'jan']


----------------------------------------------------------------------------------------------------
```

Discussion - My Solution

My solution makes use of some of the powerful tools NLTK has to offer. If I had more time I would like to use the word_lemmatize function in conjunction with another stemmer, like the Porter Stemming Algorithm which shortens words down even further (a word like ‘cry’ or ‘cries’ becomes ‘cri’) as this would save memory for larger words in larger corpora.

As an extension to the assignment, I would like to consider indexing the words and moving onto indexing by storing multiple indexed documents and comparing / querying them so that they can be used in an actual retrieval model.

Appendix 1 : Stopwords

Used NLTK.CORPUS stopwords. 153 words.

{'shouldn', "you'll", 'yourselves', 'because', 'all', 'ourselves', 'some', "should've", "hadn't", 'before', 'weren', 'd', 'when', 'have', 'too', 'such', "didn't", 'why', 'now', 'that', 'did', 'what', 'be', 'over', 'was', 'against', 'doesn', 'those', "mustn't", "weren't", "doesn't", 'himself', "haven't", 'won', 'is', 'above', 't', 'where', 'out', 'most', 'shan', 'this', 'further', 'm', "you've", 'any', 'a', "you're", 'off', "shouldn't", 'not', "aren't", 'up', 'then', "won't", 'which', 's', 'each', 'y', 'so', 'him', 'of', 'to', 'myself', 'through', "you'd", 'into', 'their', 'as', 'should', 'o', 'who', 'mustn', 'didn', 'and', 'herself', 'they', 'below', 'yourself', 'wasn', 'an', 'on', "it's", 'these', 'here', 'am', 'ain', 'very', "wouldn't", 'being', 'we', "couldn't", 'you', 'themselves', 'has', 'itself', 'after', 'hasn', 'me', 'it', 'at', 'or', 'isn', 'aren', 'hadn', 'haven', 're', 'he', 'needn', 'during', 'with', 'while', 've', 'how', 'about', 'had', "she's", 'ours', 'does', 'once', 'there', 'her', 'its', 'do', 'will', 'doing', 'were', 'i', 'just', 'she', 'in', 'both', "shan't", 'my', 'same', 'been', 'll', 'your', 'only', 'having', 'from', 'the', "don't", 'couldn', 'few', 'own', 'are', 'nor', 'can', "hasn't", "mightn't", 'but', 'mightn', 'than', 'by', 'more', 'don', 'if', 'hers', 'under', 'yours', 'theirs', 'no', "needn't", 'them', "wasn't", 'his', 'until', 'again', 'wouldn', 'ma', "isn't", 'our', 'whom', 'for', 'other', 'between', "that'll", 'down'}

Appendix 2: Output of Each Stage for URL: http://csee.essex.ac.uk/staff/udo/index.html

Output_The Udo.txt

Appendix 3: Output of Each Stage for URL: http://irsg.bcs.org/ksjaward.php

Output_KSJ
Award.txt