I tried first just using 2 implementations of Rake but results weren’t that good:

|  |  |
| --- | --- |
| Rake\_nltk with max\_mords=2 | Python-rake with max\_words=2 |
| [(4.0, 'w height'),  (4.0, 'valid --'),  (4.0, 'useful combination'),  (4.0, 'typename variablename'),  (4.0, 'typename )•'),  (4.0, 'thereby relieving'),  (4.0, 'suspend animations'),  (4.0, 'superclass view'),  (4.0, 'start animations'),  (4.0, 'sstill allocated'),  (4.0, 'square brackets'),  (4.0, 'source codedirectly'),  (4.0, 'size within'),  (4.0, 'simple applicationconsider'),  (4.0, 'shift operator'),  (4.0, 'sections ofcode'),  (4.0, 'runs within'),  (4.0, 'rights reserved'),  (4.0, 'restrictions .•'),  (4.0, 'remains allocateduntil'),  (4.0, 'put simply'),  (4.0, 'programming errors'),  (4.0, 'primitive types'),  (4.0, 'port well'),  (4.0, 'pointer types'),  (4.0, 'plus operator'),  (4.0, 'optional within'),  (4.0, 'opening /\*'),  (4.0, 'nothing points'),  (4.0, 'menu bar'),  (4.0, 'lowest priority'),  (4.0, 'longer limited'),  (4.0, 'link phase'),  (4.0, 'latest versions'),  (4.0, 'interesting observation'),  (4.0, 'initial context'),  (4.0, 'implied assignments'),  (4.0, 'ignore messages'),  (4.0, 'hostplatform ).'),  (4.0, 'harm hardware'),  (4.0, 'hardware environments'),  (4.0, 'hard disks'),  (4.0, 'gui components'),  (4.0, 'garbage collector'),  (4.0, 'garbage collection'),  (4.0, 'first element'),  (4.0, 'final clean'),  (4.0, 'extend right'),  (4.0, 'existing portions'),  (4.0, 'every platform'),  (4.0, 'easily find'),  (4.0, 'easily comment'),  (4.0, 'dispose ()'),  (4.0, 'demand atruntime'),  (4.0, 'declaration defines'),  (4.0, 'deallocated ).'),  (4.0, 'current objectsuper'),  (4.0, 'copysemantics ."'),  (4.0, 'continue apply'),  (4.0, 'completely specified'),  (4.0, 'compilation unit'),  (4.0, 'closing \*/'),  (4.0, 'clone ();'),  (4.0, 'classthat contains'),  (4.0, 'buttonsbutton []'),  (4.0, 'buttonsbutton \*\*'),  (4.0, 'becausethey handle'),  (4.0, 'always signed'),  (4.0, 'address arithmetic'),  …... | [('{case int-const-expr', 11.735),  ('stat1case int-const-expr', 11.735),  ('already-existing portions', 9.0),  ('upper-left corner', 9.0),  ('/modules/javabasics/javabasics', 9.0),  ('//depot/main/src/', 7.6),  ('c-style comments', 7.1),  ('user-defined parameters', 6.666666666666667),  ('machine-independentbyte code', 6.65),  ('five-element array', 6.588235294117647),  ('variable-length strings', 6.5),  ('<applet code=appletname', 6.0346153846153845),  ('boolean-expr', 4.208333333333333),  ('rights reserved', 4.0),  ('operating systems', 4.0),  ('hardware environments', 4.0),  ('broad market', 4.0),  ('longer limited', 4.0),  ('source codedirectly', 4.0),  ('latest versions', 4.0),  ('corrupt/steal', 4.0),  ('harm hardware', 4.0),  ('hard disks', 4.0),  ('features designed', 4.0),  ('compilation unit', 4.0),  ('acompilation unit', 4.0),  ('initial context', 4.0),  ('easily find', 4.0),  ('link phase', 4.0),  ('diagram shows', 4.0),  ('non-publicclass', 4.0),  ('demand atruntime', 4.0),  ('simple applicationconsider', 4.0),  ('trivial application', 4.0),  ('gui components', 4.0),  ('contained components', 4.0),  ('becausethey handle', 4.0),  ('final clean-', 4.0),  ('menu bar', 4.0),  ('ignore messages', 4.0),  ('simple appletthe', 4.0),  ('yourbrowser window', 4.0),  ('c-stylecomments', 4.0),  ('interesting observation', 4.0),  ('easily comment', 4.0),  ('sections ofcode', 4.0),  ('document comment', 4.0),  ('simple syntax', 4.0),  ('primitive types', 4.0),  ('objectsa simple', 4.0),  ('struct definition', 4.0),  ('garbage collection', 4.0),  ('square brackets', 4.0),  ('constantsvariables modified', 4.0),  ('namelistqualified-idinteger', 4.0),  ('char literal', 4.0),  ('current objectsuper', 4.0),  ('superclass view', 4.0),  ('lowest priority', 4.0),  ('table shows', 4.0),  ('continue apply', 4.0),  ('implied assignments', 4.0),  ('pointer types', 4.0),  ('address arithmetic', 4.0),  ('equivalentcode fragments', 4.0),  ('garbage collector', 4.0),  ('sstill allocated', 4.0),  ('ill effects', 4.0),  ('remains allocateduntil', 4.0),  ('put simply', 4.0),  ('programming errors', 4.0),  ('java programs', 3.9117647058823533),  ('java compilation', 3.9117647058823533),  ('import java', 3.9117647058823533),  ('java strings', 3.9117647058823533),  ….. |

The python-rake implementation already seemed better then the other one, maybe mostly because of it’s default stop list. Without maximum words, the rake-nltk one had very bad results so going further I used only the python-rake package.

So next my plan was to first try and see how much better would results be if I added a few more stop-words related to the code.

So I added some of the Java keywords, deleted some special characters.

Also tried to delete single line comments, assignments and code between brackets and curly braces.

After that resuls were much better:

|  |  |
| --- | --- |
| No paramaters | maxWords=2 |
| [('namelist qualified id integer max\_value', 20.933333333333334),  ('creates scoped symbol startingclassname main', 20.833333333333336),  ('object general expressions item examples', 18.333333333333332),  ('pre existing main window applets', 17.8),  ('runtime exception called arrayindexoutofboundsexception accessing',  17.61111111111111),  ('definition requires exception handling code', 17.583333333333332),  ('automatic garbage collector deallocates memory', 17.523809523809526),  ('platform window manager function calls', 16.933333333333334),  ('automatically invoke predefined lifecycle methods', 16.8),  ('page intentionally left blank', 16.0),  ('dlls dynamically loadable libraries', 14.5),  ('potential memory leakage problems', 14.333333333333334),  ('burden system resources unnecessarily', 13.875),  ('runtime nullpointerexception initializers variables', 13.875),  ('32 bit integer copy assignment', 13.5),  ('address arithmetic java variables', 13.4375),  ('common statements statement examples', 13.399999999999999),  ('constant expressions item examples', 13.133333333333333),  ('multi dimensional java arrays', 12.991071428571429),  …. | [('section include', 4.0),  ('operating systems', 4.0),  ('broad market', 4.0),  ('longer limited', 4.0),  ('rights reserved', 4.0),  ('nonportable representation', 4.0),  ('latest versions', 4.0),  ('corrupt/steal', 4.0),  ('harm hardware', 4.0),  ('hard disks', 4.0),  ('main lines', 4.0),  ('interpreter level', 4.0),  ('features designed', 4.0),  ('avoid crashes', 4.0),  ('compilation unit', 4.0),  ('initial context', 4.0),  ('www programming', 4.0),  ('major differences', 4.0),  ('publicly visible', 4.0),  ('easily find', 4.0),  ('loader searches', 4.0),  ('directories listed', 4.0),  ('link phase', 4.0),  ('diagram shows', 4.0),  ('execution sequence', 4.0),  ('existing portions', 4.0),  ('trivial application', 4.0),  ('contained components', 4.0),  ('suspend animations', 4.0),  ('system resources', 4.0),  ('frame removes', 4.0),  ('menu bar', 4.0),  ('paint graphics', 4.0),  ('ignore messages', 4.0),  ('interesting observation', 4.0),  ….. |

But still, I think playing around a bit with the parameters got me the list below.

I tried some online keyword extractors on the text and it seems we actually got better results like this.

[('java runtime system', 8.132229832572298),

('boolean expr stat1', 8.0),

('java primitive types', 7.503261578604045),

('button bfoo button', 7.266666666666666),

('applet code appletname', 6.833333333333333),

('java applet applet', 6.479452054794521),

('rights reserved', 5.333333333333333),

('java strings', 5.079452054794521),

('primitive types', 5.023809523809524),

('java programs', 4.812785388127854),

('garbage collection', 4.7),

('java objects', 4.646118721461187),

('predefined methods', 4.527777777777778),

('java applications', 4.3683409436834095),

('java language', 4.346118721461187),

('www browser', 4.318181818181818),

('style comments', 4.309523809523809),

('button object', 4.008333333333333),

('compilation unit', 4.0),

('allocates memory', 3.892857142857143),

('array object', 3.875),

('allocate memory', 3.767857142857143),

('standard output', 3.666666666666667),

('typename variablename', 3.666666666666667),

('make array', 3.666666666666667),

('string literal', 3.5142857142857142),

('parameter definitions', 3.3333333333333335),

('java', 2.4794520547945207),

('arrays', 2.4285714285714284),

('statements', 2.4),

('types', 2.357142857142857),

('equivalent', 2.3333333333333335),

('type', 2.3333333333333335),

('program', 2.25),

('makes', 2.25),

('objects', 2.1666666666666665),

('comments', 2.1666666666666665),

('expr', 2.1666666666666665),

('memory', 2.142857142857143),

('button', 2.1333333333333333),

('data', 2.111111111111111),

('defined', 2.0),

('applet', 2.0),

('/applet', 2.0),

('test', 2.0),

('array', 2.0),

('stack', 2.0),

('method', 1.9166666666666667),

('applications', 1.8888888888888888),

('file', 1.8888888888888888),

('object', 1.875),

('language', 1.8666666666666667),

('code', 1.8333333333333333),

('browser', 1.8181818181818181),

('length', 1.8),

('methods', 1.7777777777777777),

('instance', 1.75),

('operators', 1.75),

('string', 1.7142857142857142),

('classes', 1.6666666666666667),

('write', 1.6666666666666667),

('reside', 1.6666666666666667),

('allocate', 1.625),

('applets', 1.4444444444444444),

('called', 1.375),

('definition', 1.3333333333333333),

('dispose', 1.3333333333333333),

('refer', 1.3333333333333333),

('convert', 1.3333333333333333),

('library', 1.25),

('init', 1.25),

('reference', 1.1666666666666667),

('operator', 1.125),

('secure', 1.0),

('robust', 1.0),

('executed', 1.0),

('part', 1.0),

('effect', 1.0),

('contrast', 1.0),

('state', 1.0),

('subclass', 1.0),

('appletviewer', 1.0),

('creates', 1.0),

('initialized', 1.0),

('buttons', 1.0),

('added', 1.0),

('drawn', 1.0),

('trivialapplet', 1.0),

('similar', 1.0),

('form', 1.0),

('width', 1.0),

('text', 1.0),

('heap', 1.0),

('addition', 1.0),

('pointers', 1.0),

('elements', 1.0),

('5 ints', 1.0),

('references', 1.0),

('assigned', 1.0),

('delimited', 1.0),

('passed', 1.0),

('define', 1.0),

('equal', 1.0),

('comparison', 1.0),

('illustrate', 1.0),

('push', 1.0),

('function', 1.0),

('time', 1.0)]