1 Best Friends

1.1 Tasks

1) Compute the pointwise mutual information for all the possible word pairs appearing consecutively in the data, disregarding pairs in which one or both words appear less than 10 times in the corpus, and sort the results from the best to the worst (did you get any negative values? Why?) Tabulate the results, and show the best 20 pairs for both data sets.

2) Do the same now but for distant words, i.e. words which are at least 1 word apart, but not farther than 50 words (both directions). Again, tabulate the results, and show the best 20 pairs for both data sets.

1.2 Solution

1) Did you get any negative values? Why?

   The worst pointwise mutual information (PMI) was -8.79 for words "the" and "," in english text and -8.48 for words "se" and "&num" in czech text. PMI is computed as:

   $$I'(a,b) = \log_2(p(a,b)/(p(a)p(b))).$$

   We can see that the if the PMI is negative, then the inside of the log is less than 1. It can be achieved when the joint probability is less than the expected probabillity. We can interpret this property of PMI like this. If the joint probability is so low that it is smaller than expected probability of randomly shuffled words, it means that the given words "do not like each other".

   This can be caused most often by a gramatical or syntax error in the text because in case of error we get words that are not in expected collocation (expected order). In our case we got for example bigram "the ," and ", ." which are clearly syntax and grammatical errors.

English word pair	Pointwise mutual information	Czech word pair	Pointwise mutual information
La, Plata	14.169	Hamburger, SV	14.289
Asa, Gray	14.032	Los, Angeles	14.062
Fritz, Muller	13.362	Johna, Newcomba	13.763
worth, while	13.333	Č., Budějovice	13.634
faced, tumbler	13.262	série, ATP	13.469
lowly, organised	13.217	turnajové, série	13.434
Malay, Archipelago	13.110	Tomáš, Ježek	13.429
shoulder, stripe	13.054	Lidové, noviny	13.330
Great, Britain	12.915	Lidových, novin	13.271
United, States	12.847	veřejného, mínění	13.062
English, carrier	12.526	teplota, minus	12.982
specially, endowed	12.402	Ján, Čarnogurský	12.956
Sir, J	12.377	jaderné, zbraně	12.956
branched, off	12.377	Milan, Máčala	12.898
de, Candolle	12.362	lidských, práv	12.863
mental, qualities	12.362	společném, státě	12.708
Galapagos, Archipelago	12.345	akciových, společností	12.692
red, clover	12.324	Pohár, UEFA	12.625
self, fertilisation	12.317	privatizačních, projektů	12.616
systematic, affinity	12.252	George, Bushe	12.603

Table 1: Best 20 pairs of words with their pointwise mutual informatio in czech and english texts.

2) We did same with distant words:

English word pair	Pointwise mutual information	Czech word pair	Pointwise mutual information
floated, dried	15.350	výher, výher	16.470
dried, floated	15.350	žel, žel	15.750
dried, dried	14.918	13h, 13h	15.502
floated, floated	14.807	Sandžaku, Sandžaku	15.441
dried, germinated	14.780	Petrof, Petrof	15.400
germinated, dried	14.780	Bělehrad, Benfica	15.374
heath, heath	14.675	Benfica, Bělehrad	15.374
floated, germinated	14.649	CIA, CIA	15.119
germinated, floated	14.649	IFS, IFS	15.052
clover, clover	14.641	13h, zataženo	14.984
eastern, Pacific	14.501	zataženo, 13h	14.984
Pacific, eastern	14.501	Atény, Benfica	14.956
avicularia, vibracula	14.469	Benfica, Atény	14.956
vibracula, avicularia	14.469	výher, IV	14.956
cave, cave	14.432	IV, výher	14.956
dried, days	14.432	13h, skoro	14.937
days, dried	14.432	skoro, 13h	14.937
metamorphosed, metamorphosed	14.432	39, 39	14.915
heads, heads	14.420	Bělehrad, Atény	14.882
stripe, shoulder	14.416	Atény, Bělehrad	14.882

Table 2: Best 20 pairs of distant words (at least 1 word apart, but not farther than 50 words) with their pointwise mutual informatio in czech and english texts.

2 Word Classes

2.1 Tasks

1) Compute a full class hierarchy of words using the first 8,000 words of those data, and only for words occurring 10 times or more (use the same setting for both languages). Ignore the other words for building the classes, but keep them in the data for the bigram counts. For details on the algorithm, use the Brown et al. paper distributed in the class; some formulas are wrong, however, so please see the corrections on the web (Class 12, formulas for Trick #4). Note the history of the merges, and attach it to your homework.

2) Now run the same algorithm again, but stop when reaching 15 classes. Print out all the members of your 15 classes and attach them too.

2.2 Solution

1) In english text we have got 112 classes out of 1662 different words in the text. We assigned an id to each class in the text. If we have these id's of the initial classes:
(on,0), (.,3), (as,7), („,9), (I,10), (much,12), (with,14), (certain,15), (facts,16), (in,17), (the,18), (of,20), (and,26), (to,30), (that,33), (will,36), (be,37), (this,41), (some,45), (species,48), (it,51), (has,52), (been,53), (by,55), (one,56), (our,57), (my,61), (me,65), (all,76), (which,78), (could,79), (have,81), (any,82), (subject,91), (short,94), (;,96), (these,97), (a,101), (:,106), (from,107), (same,112), (may,115), (for,117), (not,124), (is,129), ((,131), (),133), (nearly,134), (but,136), (many,138), (more,139), (do,150), (at,160), (In,164), (would,170), (must,197), (cannot,200), (several,204), (can,221), (only,222), (most,225), (cases,226), (than,230), (often,246), (each,261), (very,272), (their,302), (other,306), (such,307), (varieties,315), (even,317), (if,318), (how,323), (so,327), (structure,330), (conditions,339), (variation,344), (we,347), (shall,348), (see,349), (its,356), (under,365), (case,368), (distinct,387), (or,389), (It,394), (domesticated,408), (animals,409), (plants,411), (they,434), (what,450), (great,452), (slight,457), (state,460), (nature,461), (long,470), (are,475), (individuals,496), (there,500), (manner,508), (less,531), (an,555), (when,566), (believe,613), (differ,657), (different,664), (domestic,669), (The,727), (races,780), (between,836), (wild,993), (breeds,1182)

We get this history of merges:
(368, 91), (115, 200), (496, 330), (394, 500), (317, 531), (344, 461), (348, 349), (457, 94), (15, 387), (307, 508), (368, 460), (657, 450), (226, 133), (222, 65), (197, 221), (323, 134), (408, 669), (315, 780), (318, 566), (230, 131), (150, 613), (61, 452), (36, 79), (496, 344), (16, 411), (106, 836), (356,

664), (164, 727), (339, 1182), (457, 470), (97, 261), (115, 170), (225, 138), (246, 272), (394, 434), (555, 82), (657, 12), (150, 348), (139, 317), (15, 993), (327, 136), (226, 409), (307, 230), (115, 197), (323, 76), (56, 306), (339, 368), (160, 222), (57, 61), (97, 45), (394, 347), (356, 302), (315, 48), (106, 318), (225, 204), (226, 16), (117, 365), (394, 51), (15, 408), (246, 657), (115, 36), (14, 327), (41, 555), (112, 56), (139, 457), (496, 339), (323, 124), (106, 7), (129, 52), (160, 0), (307, 389), (150, 53), (107, 55), (226, 315), (97, 356), (10, 394), (225, 57), (41, 78), (117, 164), (475, 129), (323, 246), (15, 112), (150, 81), (14, 33), (226, 496), (97, 101), (139, 307), (107, 160), (150, 37), (14, 106), (117, 96), (475, 115), (225, 15), (323, 41), (17, 107), (139, 26), (97, 323), (30, 117), (475, 150), (225, 226), (139, 14), (475, 10), (17, 20), (97, 18), (30, 3), (9, 30), (139, 475), (97, 225), (9, 17), (97, 139), (97, 9)

Where each pair of classes is merged into the class on the left.

In czech text we have got 61 classes out of 3685 different words in the text. Id's of initial classes are:
(.,0), (:,2), ((,12), („,14), (),16), (-,17), (Na,18), (byl,22), (včera,31), (i,35), (o,36), (k,41), (se,45), (a,52), (zákona,59), (do,63), (na,79), (jeho,93), (s,98), (který,108), (od,111), (v,119), (ve,130), (po,134), (že,138), (",141), (aby,154), (bude,160), (J,165), (z,167), (ze,171), (V,173), (ale,198), (to,200), (pro,212), (by,248), (být,256), (při,350), (mezi,366), (ČSFR,371), (jsou,387), (budou,390), (už,441), (za,475), (listopadu,508), (státu,521), (je,536), (pouze,545), (které,549), (nás,572), (u,593), (musí,660), (?,710), (NATO,776), (před,786), (bylo,811), (si,874), (však,912), (jako,1231), (&slash;,2012), (OKD,2152)

History of merges:
(2152, 508), (549, 108), (165, 521), (660, 160), (198, 154), (572, 811), (874, 441), (545, 256), (912, 366), (2012, 776), (171, 93), (387, 22), (786, 710), (1231, 390), (18, 2152), (59, 350), (593, 475), (165, 371), (549, 198), (874, 31), (572, 111), (660, 912), (134, 212), (18, 173), (545, 387), (41, 171), (130, 2012), (786, 1231), (549, 138), (59, 200), (572, 248), (63, 593), (35, 874), (165, 16), (536, 2), (167, 134), (660, 786), (545, 12), (98, 130), (41, 59), (18, 141), (572, 35), (63, 536), (545, 17), (165, 98), (167, 36), (549, 660), (41, 45), (572, 79), (165, 18), (167, 63), (545, 119), (549, 52), (41, 572), (167, 545), (549, 165), (41, 14), (167, 0), (549, 41), (549, 167)

2) All 15 classes of the english text are:
97: 'these', 'each', 'some', 'its', 'different', 'their', 'a', 'how', 'nearly', 'all', 'not', 'often', 'very', 'differ', 'what', 'much', 'an', 'any', 'this', 'which'
139: 'more', 'even', 'less', 'slight', 'short', 'long', 'such', 'manner', 'than', '(', 'or', 'and'
475: 'are', 'is', 'has', 'may', 'cannot', 'would', 'must', 'can', 'will', 'could'
225: 'most', 'many', 'several', 'our', 'my', 'great', 'other', 'one', 'same', 'wild', 'certain', 'distinct', 'domesticated', 'domestic'
226: 'cases', ')', 'animals', 'plants', 'facts', 'varieties', 'races', 'species', 'individuals', 'structure', 'variation', 'nature', 'conditions', 'breeds', 'case', 'subject', 'state'
394: 'It', 'there', 'they', 'we', 'it', 'I'
9: ','
106: ':', 'between', 'if', 'when', 'as', 'so', 'but', 'with', 'that'
107: 'from', 'by', 'at', 'me', 'only', 'on', 'in'
150: 'do', 'believe', 'shall', 'see', 'been', 'have', 'be'
117: 'for', 'under', 'In', 'The', ';'
18: 'the'
20: 'of'
30: 'to'
3: '.'

All 15 classes of the czech text are:
2: ':', 'je', 'u', 'za', 'do'
173: 'V', 'listopadu', 'OKD', 'Na', '"'
93: 'jeho', 'ze', 'k', 'při', 'zákona', 'to'
572: 'nás', 'bylo', 'od', 'by', 'si', 'už', 'včera', 'i'
45: 'se'
0: '.'
22: 'byl', 'jsou', 'být', 'pouze', '(', '-'
130: 've', '&slash;', 'NATO', 's', 'J', 'státu', 'ČSFR', ')'
212: 'pro', 'po', 'z', 'o'
660: 'musí', 'bude', 'však', 'mezi', 'před', '?', 'jako', 'budou'

108: ’který’, ’které’, ’ale’, ’aby’, ’že’
79: ’na’
14: ’,’
119: ’v’
52: ’a’

3 Tag Classes

3.1 Tasks

Use the same original data as above, but this time, you will compute the classes for tags (the strings after slashes).

1) Compute tag classes for all tags appearing 5 times or more in the data. Use as much data as time allows. You will be graded relative to the other student’s results. Again, note the full history of merges, and attach it to your homework. Pick three interesting classes as the algorithm goes (English data only; Czech optional), and comment on them (why you think you see those tags there together (or not), etc.).

3.2 Solution

1) In english text we have got 36 classes out of 36 different tags. Id’s of initial classes are:
(IN,0), (NN,1), (NNP,2), (.,3), („,4), (PRP,5), (VBD,6), (JJ,7), (NNS,8), (DT,9), (VBG,10), (CC,11), (TO,12), (MD,13), (VB,14), (VBN,15), (VBZ,16), (CD,17), (PRP$,18), (JJS,19), (RB,20), (WDT,21), (VBP,22), (:,23), ((,24), (SYM,25), (JJR,26), (WP,27), (RBS,28), (WRB,29), (EX,30), (",31), (NNPS,32), (FW,33), (RBR,34), (WP$,35)

   History of merges:
(34, 35), (34, 26), (32, 25), (5, 30), (33, 2), (3, 24), (27, 31), (19, 7), (34, 28), (27, 29), (18, 9), (17, 19), (32, 33), (21, 27), (6, 15), (4, 23), (16, 22), (34, 17), (10, 21), (8, 1), (4, 3), (11, 10), (16, 13), (12, 0), (32, 11), (6, 20), (6, 14), (32, 5), (6, 16), (6, 32), (6, 4), (34, 18), (6, 12), (6, 34), (6, 8)

   - As can be seen, the first two tags that are merged are RBR and WP$.
     Words with RBR tag are:
       earlier, oftener, More,and sooner.
     Word with WP$ tag is:
       whose
     In text with 221098 words, the WP$ tag occured 47 times and the RBR tag occured 15 times. This is fairly low number of occurences and it can be the main reason why the loss of mutual information by merging those two classes was the lowest one.
   - Another interesting class is class containing these tags: ’,’, ’:’, ’.’, and ’(’. The comma, colon, dot and brackets are not words in the common sense, but are used to structure sentences into larger units. The algorithm discovered this similarity and connected them to one class.
     Also the fact that the pairs (’.’,’(’) and (’,’,’:’) were connected together first and then the final class was created from these two pairs. From my experience I know, that ’,’ and ’:’ work similarly as a separator (or mark) of the side (subordinate) sentence. The dot and bracket on the other side work most often as a separator of two main sentences.
   - Another interesting merge of classes is merge of JJS and JJ.
     Words tagged with JJS are for example:
       best, oldest, commonest, humblest, slightest, plainest, closest, merest, finest, etc.
     Words tagged with JJ are for example:
       synthetic, brief, emphatic, Serious, aware, exciting, nasal, abundant, mental, etc.
     There are in total 60 different words with JJS tag occuring 487 times in total and 130 different words with JJ tag occuring 19808 times. The JJS are probably adjectives in superlative form and JJ are adjectives in normal form without suffix "-er" or "-est". This merge is in my opinion expected as the adjective as whole appear in similar combinations with other parts of speech. There is in my opinion only slight difference between surrounding of adjectives in normal, comparative, and superlative form.