Prob 1.

A.

Y	P(H=1|Y)	P(U=1|Y)	P(L=1|Y)	Prior P(Y)
1	0.667	0.833	0.333	0.5
0	0.333	0.5	0.333	0.5

B.

M: $$P(Y=1|H=0,U=1,L=0) = \frac{P(H=0,U=1,L=0|Y=1)P(Y=1)}{P(H=0,U=1,L=0)}$$

$$= \frac{P(H=0|Y=1)P(U=1|Y=1)P(L=0|Y=1)P(Y=1)}{P(H=0,U=1,L=0)}$$

$$= \frac{P(H=0|Y=1)P(U=1|Y=1)P(L=0|Y=1)P(Y=1)}{P(H=0,U=1,L=0|Y=1)P(Y=1)P(H=0,U=1,L=0|Y=0)P(Y=0)}$$

$$= \frac{0.333 * 0.833 * 0.667 * 0.5}{0.333 * 0.833 * 0.667 * 0.5 + 0.667 * 0.5 * 0.667 * 0.5}$$

$$= 0.455$$

M: $$P(Y=0|H=0,U=1,L=0) = \frac{P(H=0|Y=0)P(U=1|Y=0)P(L=0|Y=0)P(Y=0)}{P(H=0,U=1,L=0)}$$

$$= \frac{0.5 * 0.5 * 0.667 * 0.5}{0.333 * 0.833 * 0.667 * 0.5 + 0.667 * 0.5 * 0.667 * 0.5} = 0.545$$

$\therefore$ $P(Y=0|H=0,U=1,L=0) > P(Y=1|H=0,U=1,L=0)$

The result show this email is not a spam.

C. $P(Y=1|H=0,U=1,L=0) = 0.5$

$P(Y=0|H=0,U=1,L=0) = 0.5$

No,the result will be different. The reason Is if we get the probability directly from instances, we only consider 2 of them. For method B, we have another assumption is independence of features. It will not maintain here.

D. $P(Y=1) + P(Y=0) = 1$

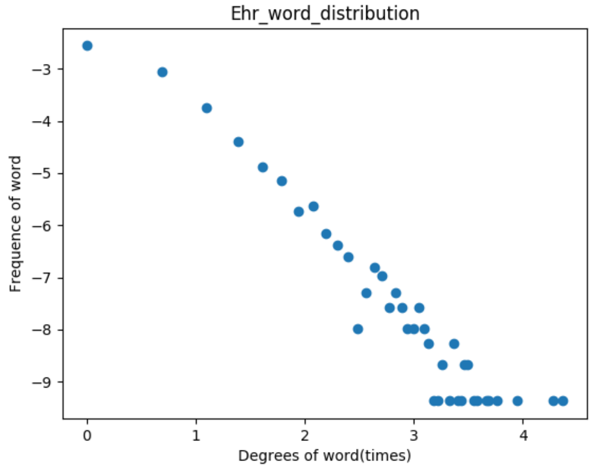
E. We want this email to be a spam one. We could change the first row Y=1, from H=1 to H = 0 so that the $P(Y=1|H=0,U=1,L=1) > P(Y=0|H=0,U=1,L=0)$

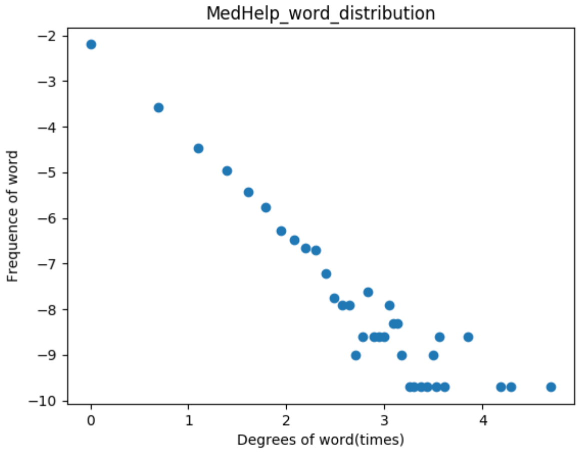
F. If we would get the direct value. We need 14 values. For Y=1 and Y=0, each situation needs 2^3-1=7 values since we have 3 variables.

G: For example, some email with url in it, the length of email will naturally more longer.

Prob 2

1. Plot the distributions on a log-log scale x-axis - word frequency (number of times a word appears in the collection); y-axis - proportion of words with this frequency.
   The x and y axis are already log scale
   A. Both are power low distributions.
   B. They are similar.





2. Properties.
   a. Frequency of stop word: EHR 0.354 < Med-Help 0.565
      Med-Help users usually use oral English, so there will be more stop word in Med-Help collocation.
   b. Percentage of capital letters: EHR: 0.063 > Med-Help: 0.033
      EHRs are more written officially using terminology such as medicine name or treatment name usually begin with capital letters.
   c. Average number of characters per word: EHR: 5.025 > MedHelp:4.147
      Terminologies in EHR will make contributions for average length of words.

   d. Percentage of nouns, adjectives, verbs, adverbs, and pronouns
      EHR: NOUN: 0.38, ADJ: 0.12, VERB: 0.14, ADV: 0.033, PRON: 0.038
      Med-Help: NOUN: 0.21, ADJ: 0.08, VERB: 0.28, ADV: 0.079, PRON: 0.117
      EHR has more Noun, ADJ and Med-Help has more Verb, ADV, pron.
      The reason is still the core difference between professional written English and Spoken English in forum.
   e. The top 10 nouns, top 10 verbs, and top 10 adjectives

EHR:

Noun['pain', 'patient', 'No', 'history', 'home', 'ED', 'days', 'Pt', 'fibrillation', 'weeks']

Verb['was', 'given', 'had', 'found', 'have', 'presented', 'be', 'has', 'denies', 'showed']

ADJ['abdominal', 'atrial', 'old', 'right', 'positive', 'recent', 'pulmonary', 'negative', 'chest', 'unresponsive']

Med-Help:

Noun['time', 'people', 'i', 'weeks', 'days', 'things', 'pain', 'doctor', 'years', 'help']

Verb['have', 'be', 'is', 'know', "'m", 'are', 'think', 'take', 'did', 'can']

ADJ['good', 'i', 'normal', 'sure', 'long', 'right', 'low', 'hard', 'new', 'bad']

The difference of written professional English and oral English in forum causes the difference of distribution of tags in collocations.
More sentence begin with word ‘I ’ is one feature of spoken English in forum. In EHR, ‘patient’ and ‘doctor’ are used to refer different persons.

3. For each doc print (top 5 words)
   I don’t use stemmer but I lower the word
   EHR
   doc0 ['frank', 'melanotic', 'nqwmi', 'quantity', 'readmitted']
   doc1 ['snap', 'hip', 'leg', 'falling', 'arthroplasty']
   doc2['hypotension', 'group', '29', 'eventually', 'lightheadedness']
   doc3['c2', 'fracture', 'fall', 'nursing', '106/42']
   doc4['status-post', 'replacement', 'atrial', 'valve', 'bioprosthetic']
   doc5['stools', 'stricture', 'transfusion', 'earlier', 'cramping']
   doc6['tests', 'quadrant', 'ascites', 'gtt', 'dark']
   doc7['reglan', 'ago', 'menstrual', 'ivf', 'shot']
   doc8['delivery', 'mother', 'gm', 'labor', '678']
   doc9['ef', 'varices', 'dysfunction', '55y/o', 'cardiomyopathy']
4. For some other that can be considered into document ranking, one thing is that we could divide all words into different categories and for each documents we could count how many words in each categories (need normalization) and represent doc into a theme vector and for query we also use categories vector to embed. The similarity between query and doc theme vector could also be weighs considered in ranking.

Prob3

A. Precision: 8/16 = 0.5 Recall = 8/10 = 0.8, F1 = 2*0.5*0.8 / (0.5+0.8) = 0.615
   MAP = [1 + 2/3 +3/5+4/6+5/10+6/11+7/14+8/16] / 10 = 0.498

B. CG@10 = 2+ 2+1+1+2 = 8
   DCG@10 = 2+ 2/log2(3) + 1/log2(5) + 1/log2(6) + 2/log2(10) = 4.68

DCG_perfect@10 = 2 + 2/log2(2) + 2/log2(3) + 2/log2(4) + 1/log2(5) + 1/log2(6) + 1/log2(7) + 1/log2(8) + 1/log2(9) + 1/log2(10) = 8.39
NDCG@10 = 4.68/8.39 = 0.55