Lucas Sipos

Problem 1:
Text: USA a.k.a. U.S.A. has one of the world\'s largest technology companies, the I.B.M. company. It\'s amazing how U.K.-based companies also thrive globally. Some hyphenated-words like well-being or state-of-the-art or problem-solving, require careful tokenization. Does "Mr. Dr. O\'Connell" believe that A.I. will change the future? You\'d think so! However, co-working and emails\' importance are undeniable. Let\'s analyze: e.g. how tokenizers work.

a) ['USA', 'a.k.a', '.', 'U.S.A.', 'has', 'one', 'of', 'the', 'world', "'s", 'largest', 'technology', 'companies', ',', 'the', 'I.B.M', '.', 'company', '.', 'It', "'s", 'amazing', 'how', 'U.K.-based', 'companies', 'also', 'thrive', 'globally', '.', 'Some', 'hyphenated-words', 'like', 'well-being', 'or', 'state-of-the-art', 'or', 'problem-solving', ',', 'require', 'careful', 'tokenization', '.', 'Does', '``', 'Mr.', 'Dr.', "O'Connell", "''", 'believe', 'that', 'A.I', '.', 'will', 'change', 'the', 'future', '?', 'You', "'d", 'think', 'so', '!', 'However', ',', 'co-working', 'and', 'emails', "'", 'importance', 'are', 'undeniable', '.', 'Let', "'s", 'analyze', ':', 'e.g', '.', 'how', 'tokenizers', 'work', '.']

Abbreviations are mostly correct, but a.k.a., I.B.M., A.I. and e.g. are incorrectly written because the last dot is separated. A suggestion would be to continue the tokanization of an ongoing abbreviation until there is a separator found (except ‘.’ but not ‘..’).
The words with single quotes in them are done correctly, even O’Connell and emails’.
The words with hyphens are consistent.
If we ignore the abbreviation problem, then the punctuation is done correctly as well.

b) ['usa', 'a.k.a', '.', 'u.s.a.', 'has', 'one', 'of', 'the', 'world', "'s", 'largest', 'technology', 'companies', ',', 'the', 'i.b.m', '.', 'company', '.', 'it', "'s", 'amazing', 'how', 'u.k.-based', 'companies', 'also', 'thrive', 'globally', '.', 'some', 'hyphenated-words', 'like', 'well-being', 'or', 'state-of-the-art', 'or', 'problem-solving', ',', 'require', 'careful', 'tokenization', '.', 'does', '``', 'mr.', 'dr.', "o'connell", "''", 'believe', 'that', 'a.i', '.', 'will', 'change', 'the', 'future', '?', 'you', "'d", 'think', 'so', '!', 'however', ',', 'co-working', 'and', 'emails', "'", 'importance', 'are', 'undeniable', '.', 'let', "'s", 'analyze', ':', 'e.g', '.', 'how', 'tokenizers', 'work', '.']

c) [('usa', 'JJ'), ('a.k.a', 'NN'), ('.', '.'), ('u.s.a.', 'NN'), ('has', 'VBZ'), ('one', 'CD'), ('of', 'IN'), ('the', 'DT'), ('world', 'NN'), ("'s", 'POS'), ('largest', 'JJS'), ('technology', 'NN'), ('companies', 'NNS'), (',', ','), ('the', 'DT'), ('i.b.m', 'NN'), ('.', '.'), ('company', 'NN'), ('.', '.'), ('it', 'PRP'), ("'s", 'VBZ'), ('amazing', 'JJ'), ('how', 'WRB'), ('u.k.-based', 'JJ'), ('companies', 'NNS'), ('also', 'RB'), ('thrive', 'VBP'), ('globally', 'RB'), ('.', '.'), ('some', 'DT'), ('hyphenated-words', 'NNS'), ('like', 'IN'), ('well-being', 'NN'), ('or', 'CC'), ('state-of-the-art', 'NN'), ('or', 'CC'), ('problem-solving', 'NN'), (',', ','), ('require', 'VB'), ('careful', 'JJ'), ('tokenization', 'NN'), ('.', '.'), ('does', 'VBZ'), ('``', '``'), ('mr.', 'VB'), ('dr.', 'NN'), ("o'connell", 'PRP'), ("''", "''"), ('believe', 'VBP'), ('that', 'IN'), ('a.i', 'NN'), ('.', '.'),

('will', 'MD'), ('change', 'VB'), ('the', 'DT'), ('future', 'NN'), ('?', '.'), ('you', 'PRP'), ("'d", 'MD'), ('think', 'VB'), ('so', 'RB'), ('!', '.'), ('however', 'RB'), (',', ','), ('co-working', 'JJ'), ('and', 'CC'), ('emails', 'NNS'), ("'", 'POS'), ('importance', 'NN'), ('are', 'VBP'), ('undeniable', 'JJ'), ('.', '.'), ('let', 'NN'), ("'s", 'POS'), ('analyze', 'NN'), (':', ':'), ('e.g', 'NN'), ('.', '.'), ('how', 'WRB'), ('tokenizers', 'JJ'), ('work', 'NN'), ('.', '.')]

List of wrong pos tags and their correct tags:
('usa', 'JJ') - NNP, ('a.k.a', 'NN') - RB, ('u.s.a.', 'NN') - NNP, ('i.b.m', 'NN') - NNP, ('mr.', 'VB') - NNP, ('dr.', 'NN') - NNP, ("o'connell", 'PRP') - NNP, ('let', 'NN') - VB, ("'s", 'POS') - PRP,('analyze', 'NN') - VB, ('e.g', 'NN') - RB, ('tokenizers', 'JJ') - NNS, ('work', 'NN') - VBP

Problem 2:
Text: The children are playing outside while the adults are discussing unpredictable topics. She ran faster than her friend. They have been studying the new technologies. People enjoy working in teams, collaborating and solving problems. Some situations call for drastic measures, but we handled them cautiously. The running man seemed determined to reach the destination on time.

a) ['the', 'children', 'are', 'play', 'outsid', 'while', 'the', 'adult', 'are', 'discuss', 'unpredict', 'topic', '.', 'she', 'ran', 'faster', 'than', 'her', 'friend', '.', 'they', 'have', 'been', 'studi', 'the', 'new', 'technolog', '.', 'peopl', 'enjoy', 'work', 'in', 'team', ',', 'collabor', 'and', 'solv', 'problem', '.', 'some', 'situat', 'call', 'for', 'drastic', 'measur', ',', 'but', 'we', 'handl', 'them', 'cautious', '.', 'the', 'run', 'man', 'seem', 'determin', 'to', 'reach', 'the', 'destin', 'on', 'time', '.']

List of oddities and the correct way:
outsid - outside, studi - study, technolog - technology, peopl - people, collabor - collaborate, solv - solve, situat - situation, measur - measure, handl - handle, destin - destination
Most of the words are incomplete, but there are some that remove unnecessary affixes, such as destination, which has been stemmed to destin and feels odd, but a normal behavior because of the ending “-ation”. Also, “children” is not lemmetiezed to “child”.

b) ['The', 'child', 'be', 'play', 'outside', 'while', 'the', 'adult', 'be', 'discuss', 'unpredictable', 'topic', '.', 'She', 'run', 'faster', 'than', 'her', 'friend', '.', 'They', 'have', 'be', 'study', 'the', 'new', 'technology', '.', 'People', 'enjoy', 'work', 'in', 'team', ',', 'collaborate', 'and', 'solve', 'problem', '.', 'Some', 'situation', 'call', 'for', 'drastic', 'measure', ',', 'but', 'we', 'handle', 'them', 'cautiously', '.', 'The', 'run', 'man', 'seem', 'determine', 'to', 'reach', 'the', 'destination', 'on', 'time', '.']

The problems are the following: ‘unpredictable’ - ‘predict’, ‘faster’ - ‘fast’, ‘cautiously’ - ‘cautious’. Also, I would point out (as said in the lecture) are/been have been converted to ‘be’.

c) After comparing the outputs of porter stemming and lemmatization, we could see that the clarity and precision of the lemmatization is better. But, I would say porter stemming does a good job, at least better than lemmatization, at converting adjectives and adverbs, as for the example that I’ve provided.