2. The word’s probability only depends on the n previous words
3. Not necessarily, a word’s probability is likely dependent on all previous words
4. Easy to estimate/compute
6. parameters: 8

the: 3/11

sun: 1/22

rises: 1/11

in: 2/11

east: 1/11

and: 1/11

sets: 1/11

west: 1/11

1. parameters: 12

a: 0

the: 3/11

from: 0

retrieval: 0

sun: 1/11

rises: 1/11

in: 1/11

BM25: 0

east: 1/11

sets: 1/11

and: 1/11

west: 1/11

2. parameters = 12^2 = 144
3. 1/144
4. 2/144
5. 0
6. 0
7. 0
8. 0
9. 0
10. 1/144
12. Most of the probabilities across all the bigrams are 0
13. The problem is still present, but the probability distribution is not as sparse
14. the problem is more sever in the bigram model because for a bigram model the size of the vocabulary becomes |v|^2
15. use shorter n-grams and unigrams. Treat shorter n-grams as prior in Bayesian estimation
17. 2 + 0 + 0 + 0 + 1 + 4 = 7
18. For all words in d, the IDF will go down. For the words not in d, the IDF will go up
19. The Euclidian distance measures the actual distance between two vectors. Cosine similarity measures the angles between two vectors and thus does not account for their magnitude’s. Euclidian distance might be preferred because it accounts for the magnitude of the vectors
20. We can add a small constant to the TF normalization formula. The parameter is called δ and should be defaulted to 1.

h.

iraqi 0.0047735187033943055

tonight 0.0046892772585990904

terrorist 0.00410487843408031

retreat 0.004077744545089247

iraq 0.004077744545089247

terrorists 0.003899971262148911

hopeful 0.00389179803252617

iraqis 0.0038282086113870955

dan 0.0037491191379243486

math 0.0035648285966880995

terror 0.0030667760602590926

competitive 0.0030543773652652964

offensive 0.0030175933446262527

isolationism 0.003016859837396887

america 0.002881194217866025

fight 0.002864591686093277

2001 0.0027479660070170776

qaeda 0.0027479660070170776

baby 0.0026412535840269476

iran 0.0025453257113326475

Based on the terms and weights in this speech, it seems this is the state of the union in 2001, likely following the 9/11 terrorist attacks.

1900s: [('interstate', 0.016797539437369634), ('corporations', 0.01600232961465208), ('islands', 0.01363183950922941), ('canal', 0.01292735457905591), ('panama', 0.01288204399027719), ('isthmus', 0.01133762669760487), ('man', 0.010851283337031342), ('1900', 0.010563456241978329), ('forest', 0.01043650594983841), ('1899', 0.010363915123243085), ('philippines', 0.009882627870322491), ('philippine', 0.009575478532892381), ('hague', 0.009357513763560409), ('conference', 0.00931745658509235), ('tariff', 0.009223234724560814), ('wageworker', 0.009171137974833505), ('1909', 0.00885201503619397), ('industrial', 0.008768498992981823), ('cable', 0.008575666905548625), ('wageworkers', 0.008474716595952154)]

1910s: [('interstate', 0.019136724656161255), ('railways', 0.015930353256625323), ('processes', 0.015228790177484184), ('eighthour', 0.014266278437393283), ('tasks', 0.012724361533874965), ('upon', 0.012331512155983227), ('railroads', 0.012014294236672239), ('unrest', 0.011813601521404898), ('porto', 0.011709582950272052), ('industrial', 0.011551565972769384), ('programme', 0.011190669184578776), ('german', 0.010767400094019351), ('debatable', 0.01046963179216696), ('woodrow', 0.010242380035256664), ('counsel', 0.01010500258709809), ('shall', 0.009814179465232824), ('1911', 0.009752558197001777), ('thing', 0.009620576362113848), ('touch', 0.009397975699506884), ('presently', 0.009368068110373006)]

1920s: [('marketing', 0.016857993904502988), ('veterans', 0.015400703068693636), ('method', 0.015209208495676581), ('reclamation', 0.014430539517056218), ('ought', 0.013296999582226382), ('agriculture', 0.013154081676656531), ('farmer', 0.0124518571027218), ('interstate', 0.01232647273751982), ('freight', 0.011827730981477652), ('problem', 0.011783824962348005), ('cent', 0.011533849868944318), ('storage', 0.011464491103941205), ('democracy', 0.011402106138058643), ('court', 0.01134916083766858), ('1920', 0.011327609484225686), ('conference', 0.011306729375687727), ('floating', 0.010912236304282381), ('economic', 0.01077772850534639), ('agricultural', 0.010723601085750125), ('cooperative', 0.010702310217963387)]

1930s: [('1933', 0.033265956894717195), ('recovery', 0.028582270513911746), ('1928', 0.0243802380768224), ('democracy', 0.0238811928283385), ('1932', 0.023143814983038854), ('1931', 0.01877589340955795), ('economic', 0.018695378933101907), ('unemployment', 0.018628403912143925), ('industrial', 0.018221095810176915), ('depression', 0.017374555702379068), ('banks', 0.017067079490417907), ('autocracy', 0.01538063273738662), ('relief', 0.015269509303188081), ('unemployed', 0.014793791803406574), ('1930', 0.01318741414541836), ('66', 0.013052470419483283), ('problems', 0.012832442074434205), ('policies', 0.012418648306975916), ('678', 0.012411461112899879), ('farm', 0.01231737256253555)]

1940s: [('1947', 0.029710853274099887), ('1946', 0.02853434464846728), ('1942', 0.026932480971111956), ('program', 0.024786131244630635), ('japanese', 0.02278813572640153), ('economic', 0.02148080151444126), ('housing', 0.02002282372194343), ('production', 0.019710398428197957), ('fighting', 0.019568541589539582), ('democratic', 0.018254714130335657), ('cannot', 0.016848673264592447), ('hitler', 0.016434919251571024), ('1945', 0.01622232616659913), ('axis', 0.01538311833821046), ('veterans', 0.014787710002442576), ('1943', 0.014612800691498564), ('1944', 0.014524511871056902), ('million', 0.014308652850507017), ('planes', 0.014099673852278356), ('today', 0.013826782431688705)]

1950s: [('program', 0.033751704743832475), ('economic', 0.029904755668768598), ('communist', 0.025288667191289922), ('programs', 0.02347390037986735), ('atomic', 0.022549891600393192), ('budget', 0.013554119454022783), ('farm', 0.013317160714324185), ('insurance', 0.013015051704948244), ('help', 0.012660441186773473), ('korea', 0.01238939287960758), ('1954', 0.012343259879003195), ('activities', 0.01231556229082954), ('federal', 0.012276266851987731), ('weapons', 0.012050250496977371), ('strength', 0.011943212127829783), ('cannot', 0.011709801058377052), ('billion', 0.011481332996556865), ('areas', 0.011280164583333026), ('objectives', 0.011097830977717937), ('achieve', 0.011094597139524622)]

1960s: [('vietnam', 0.05209676744675074), ('tonight', 0.03738480636601402), ('billion', 0.031619277126822706), ('help', 0.023898743994768858), ('communist', 0.02102038233229099), ('program', 0.019865424033602905), ('1964', 0.019475889526783693), ('programs', 0.018929265725544132), ('americans', 0.01864734706950259), ('poverty', 0.01861884010907119), ('percent', 0.017752138828137264), ('budget', 0.01769725441778147), ('million', 0.016184169036134826), ('alliance', 0.014505489105356234), ('cannot', 0.013961016356921964), ('soviet', 0.013767978565778214), ('1968', 0.013503906476826793), ('nuclear', 0.013420512983878252), ('urban', 0.013018463262997666), ('jobs', 0.012627914374318141)]

1970s: [('programs', 0.034693051853809), ('inflation', 0.03358177749610683), ('1973', 0.03211700444157631), ('americans', 0.030167987714430473), ('spending', 0.025045596093629913), ('1974', 0.02372785325785917), ('nuclear', 0.023197190367011657), ('jobs', 0.022181632650129453), ('oil', 0.021624760270794418), ('help', 0.021459000919426877), ('percent', 0.02115667592311246), ('budget', 0.02060419223045937), ('tonight', 0.020411043639859586), ('1975', 0.020117986669535758), ('goals', 0.020018614638081737), ('program', 0.01990022764910677), ('energy', 0.01975109121411512), ('goal', 0.019538470411691756), ('major', 0.018507017497941886), ('billion', 0.017975318212725147)]

1980s: [('soviet', 0.04470693581835627), ('tonight', 0.04159979698345362), ('weve', 0.036340624372206004), ('budget', 0.026701969621960232), ('im', 0.02513365431306626), ('lets', 0.02376192079156672), ('spending', 0.023076114667794273), ('americas', 0.022744161299306034), ('programs', 0.02185088358027777), ('1980', 0.019927166976310473), ('economic', 0.018644063393941115), ('help', 0.018451315853125712), ('ive', 0.018442895230115743), ('deficits', 0.017499738869848427), ('afghanistan', 0.017190100359748838), ('thats', 0.017179771473820882), ('nuclear', 0.016977778984060943), ('americans', 0.01681173174031436), ('percent', 0.0167998120523602), ('jobs', 0.01662385949910553)]

1990s: [('tonight', 0.05075159458649747), ('jobs', 0.03090291261205944), ('americans', 0.029582210123157388), ('thats', 0.029264273613086915), ('parents', 0.027204689448909726), ('children', 0.02346594917698667), ('budget', 0.023403793985320623), ('help', 0.023384118660807897), ('challenge', 0.02263480791996252), ('dont', 0.022268490683390322), ('weve', 0.021594481472633206), ('21st', 0.020613270435321143), ('get', 0.019423159417390645), ('know', 0.019112862295406083), ('kids', 0.018816759281882214), ('child', 0.018807995626985295), ('cut', 0.018445704934533364), ('lets', 0.01835161783932196), ('tell', 0.018198197982423676), ('college', 0.017690596523088183)]

2000s: [('terrorists', 0.062422185821378096), ('iraq', 0.05728735966971791), ('tonight', 0.0459201140641024), ('terror', 0.03814323599964809), ('iraqi', 0.03676398610658946), ('terrorist', 0.03437968843468896), ('al', 0.029302002710768528), ('americans', 0.028484183756049587), ('afghanistan', 0.02802784061912633), ('help', 0.02328007988071947), ('saddam', 0.02223554875926235), ('regime', 0.019873104500498697), ('iraqis', 0.019756312208032433), ('america', 0.019692373237990373), ('weapons', 0.01938456138519062), ('hussein', 0.019355777084057297), ('jobs', 0.019179880858946865), ('qaeda', 0.018701771319434218), ('homeland', 0.0184697952656494), ('fight', 0.016021984198600115)]

2010s: [('thats', 0.025609655013987286), ('jobs', 0.020738092572593302), ('businesses', 0.01397200095768506), ('weve', 0.013755717128238563), ('im', 0.01325032026995216), ('tonight', 0.012748977994271927), ('lets', 0.012674776625430616), ('cant', 0.010679165867013323), ('dont', 0.010619612110697299), ('why', 0.009442176980532221), ('get', 0.00941740956594792), ('americans', 0.009413876180788816), ('clean', 0.009341598064758275), ('innovation', 0.009071630900698667), ('theyre', 0.008794377487198132), ('college', 0.007631462912501098), ('arent', 0.007359526849967284), ('whats', 0.007117353482162579), ('trillion', 0.006850891420090906), ('kids', 0.006754827711591622)]

It is very interesting to see the patterns of terms across decades. In the 1900s and 1910s, you can see a lot of reference to the building of the Panama canal with the words “island”, “Panama”, and “Canal” being weighted very heavily. Furthermore there is a lot of weight on “railways” and “Interstate”, which likely reflects the expansion of the transportation at this time. As we move into the 1930s and 1940s, you can see the marks of the great depression with words like “recovery” and “economic” being very prevalent. As we move towards the 1960s you can see that this is when we fought the war with Vietnam, with the word “Vietnam” being the most highly weighted word in the 1960s. In the 70s, you can see our focus shift towards economics with words like ‘programs’, ‘inflation’, and ‘spending’ all being ranked very highly. The 1980s, was the height of the cold war, which is evident with the word ‘soviet’ being the most highly weighted word. In the 1990s it seems the most prevalent topic shifts back to economics and family, with words like ‘jobs’, ‘parents’, and ‘children’ being weighted highly. The 200s was the height of the war on terror, which makes sense as the highest ranked words are ‘terrorists’, ‘iraq’, and ‘terror’. The 2010s, we shift our focus back to business, and economics, with words like ‘jobs’, and ‘businesses’ being ranked very highly.