Text Reuse and Cascading Behavior

Sandeep Soni

04/16/2024

LeBron James

85 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

"LeBron" redirects here. For his son LeBron James Jr., see Bronny James. For other people with the name, see Lebrón.

LeBron Raymone James Sr. (/ləˈbrɒn/ *lə-BRON*; born December 30, 1984) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "**King James**", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals, including eight consecutive Finals appearances from 2011 to 2018.[1] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 19 times, selected to the All-NBA Team 19 times (including 13 First Team selections)[2][3] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.[4][5] As of October 2023, he is the oldest player in the NBA.[6]

James grew up playing basketball for St. Vincent–St. Mary High School in his hometown of Akron, Ohio. He was heavily touted by the national media as a future NBA superstar for his all-around scoring, passing, athleticism and playmaking abilities.[7] A prep-to-pro, he was selected by the Cleveland Cavaliers with the first overall pick of the 2003 NBA draft. Named the 2004 NBA Rookie of the Year,[8] he soon established himself as one of the league's premier players, leading the Cavaliers to their first NBA Finals appearance in 2007 and winning the NBA MVP award in 2009 and 2010.[4] After failing to win a championship with Cleveland, James left in 2010 as a free agent to join the Miami Heat;[9] this was announced in a nationally televised special titled *The Decision* and is among the most controversial free agency moves in sports history.[10]

LeBron James

James with the Los Angeles Lakers in 2022

No. 23 – Los Angeles Lakers	
Position	Small forward / power forward
League	NBA
Personal information	
Born	December 30, 1984 (age 38)

Summarize the following paragraph in two sentences:
LeBron Raymone James Sr. (/ləˈbrɒn/ lə-BRON; born December 30, 1984) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals, including eight consecutive Finals appearances from 2011 to 2018.[1] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 19 times, selected to the All-NBA Team 19 times (including 13 First Team selections)[2][3] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.[4][5] As of October 2023, he is the oldest player in the NBA.[6]

View other drafts

LeBron James is an American professional basketball player widely regarded as one of the greatest players in the history of the sport. He has won four NBA championships, four Most Valuable Player (MVP) Awards, and four Finals MVP Awards, and has been named an All-Star 19 times.

LeBron Raymone James Sr. (/ləˈbrɒn/ lə-BRON; born December 30, 1984) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals, including eight consecutive Finals appearances from 2011 to 2018.[1] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 19 times, selected to the All-NBA Team 19 times (including 13 First Team selections)[2][3] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.[4][5] As of October 2023, he is the oldest player in the NBA.[6]

LeBron James is an American professional basketball player widely regarded as one of the greatest players in the history of the sport. He has won four NBA championships, four Most Valuable Player (MVP) Awards, and four Finals MVP Awards, and has been named an All-Star 19 times.

LeBron Raymone James Sr. (/ləˈbrɒn/ lə-BRON; born December 30, 1984) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals, including eight consecutive Finals appearances from 2011 to 2018.[1] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 19 times, selected to the All-NBA Team 19 times (including 13 First Team selections)[2][3] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.[4][5] As of October 2023, he is the oldest player in the NBA.[6]

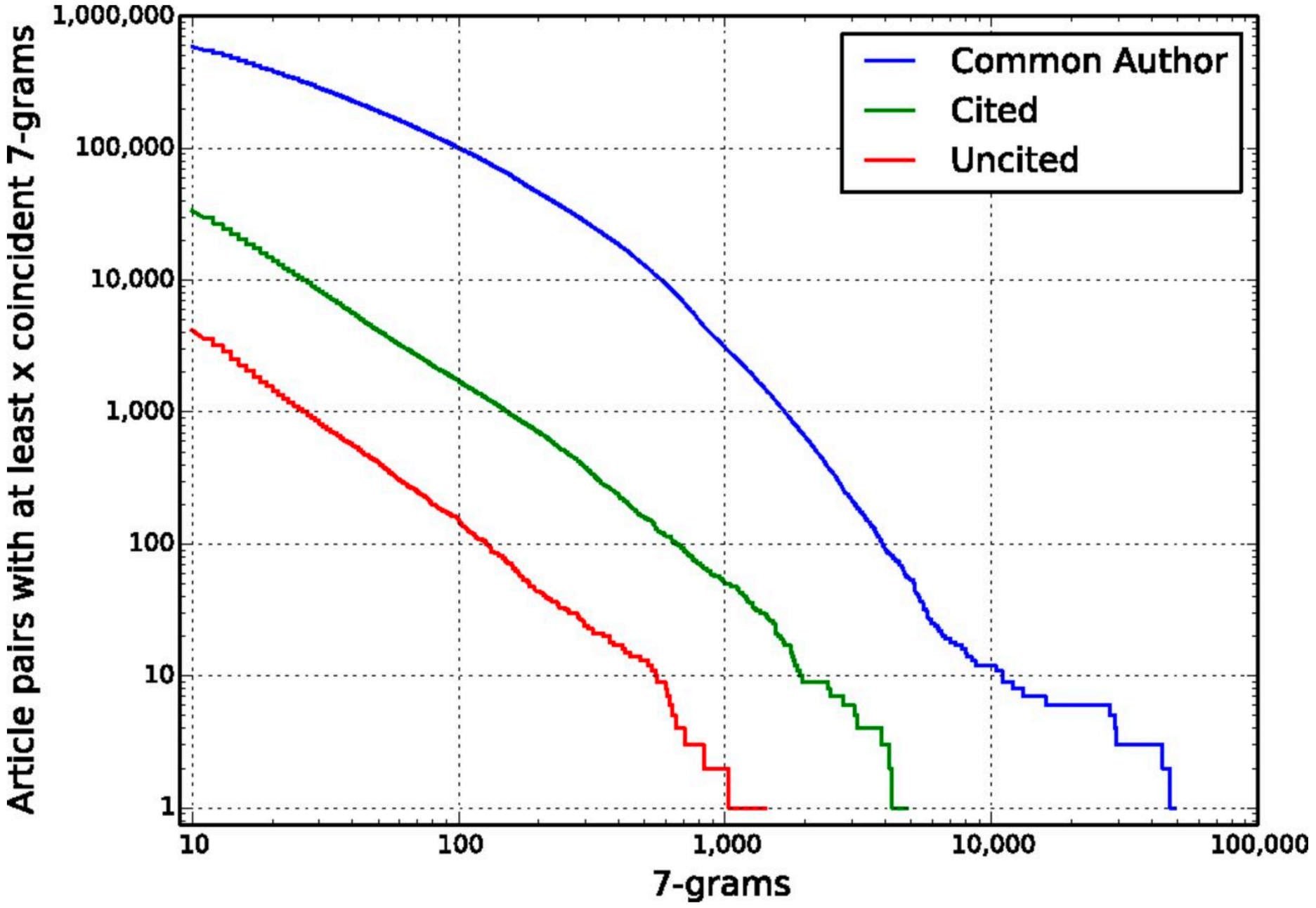
LeBron James is an American professional basketball player widely regarded as one of the greatest players in the history of the sport. He has won four NBA championships, four Most Valuable Player (MVP) Awards, and four Finals MVP Awards, and has been named an All-Star 19 times.

LeBron Raymone James Sr. (/ləˈbrɒn/ lə-BRON; born December 30, 1984) is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals, including eight consecutive Finals appearances from 2011 to 2018.[1] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 19 times, selected to the All-NBA Team 19 times (including 13 First Team selections)[2][3] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.[4][5] As of October 2023, he is the oldest player in the NBA.[6]

LeBron James is an American professional basketball player widely regarded as one of the greatest players in the history of the sport. He has won four NBA championships, four Most Valuable Player (MVP) Awards, and four Finals MVP Awards, and has been named an All-Star 19 times.

Text reuse

- As the name suggests, text reuse is a form of copying, borrowing or repeating parts of text either within a document or across documents.



Citron and Ginsparg (2014) Patterns of text reuse in a scientific corpus

Christopher Marlowe, *The Passionate Shepherd to His Love*

Come live with me and be my love,

And we will all the pleasures prove,

That Valleys, groves, hills, and fields,

Woods, or steepy mountain yields.

And we will sit upon the Rocks,

Seeing the Shepherds feed their flocks,

By shallow Rivers to whose falls

Melodious birds sing Madrigals.

And I will make thee beds of Roses

And a thousand fragrant posies,

Annie Lennox, *Live with me and be my love*

Live with me, and be my love,

And we will all the pleasures prove

By hills and valleys, dales and fields,

And all the pleasant pastures yields.

There will we sit upon the rocks,

And see the shepherds feed their flocks,

By shallow rivers, by whose falls

Melodious birds sing madrigals.

There will I make thee a bed of roses,

With a thousand fragrant posies.

Franzini, G., Franzini, E., Büchler, M. (2016) *Historical Text Reuse: What Is It?.* <http://www.etrap.eu/historical-text-re-use/>

Why quantify text reuse/duplication?

- Ability to identify duplicated text comes in handy in developing language technology

Deduplicating Training Data Makes Language Models Better

Katherine Lee[*†] **Daphne Ippolito**[*†‡] **Andrew Nystrom**[†] **Chiyuan Zhang**[†]

Douglas Eck[†] **Chris Callison-Burch**[‡] **Nicholas Carlini**[†]

Abstract

We find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data. We develop two tools that allow us to deduplicate training datasets—for example removing from C4 a single 61 word English sentence that is repeated over 60,000 times. Deduplication allows us to train models that emit memorized text ten times less frequently and require fewer training steps to achieve the same or better accuracy. We can also reduce train-test overlap, which affects over 4% of the validation set of standard datasets, thus allowing for more accurate evaluation. Code for deduplication is released at https://github.com/google-research/deduplicate-text-datasets.

We show that one particular source of bias, duplicated training examples, is pervasive: all four common NLP datasets we studied contained duplicates. Additionally, all four corresponding validation sets contained text duplicated in the training set. While naive deduplication is straightforward (and the datasets we consider already perform some naive form of deduplication), performing thorough deduplication at scale is both computationally challenging and requires sophisticated techniques.

We propose two scalable techniques to detect and remove duplicated training data. *Exact* substring matching identifies verbatim strings that are repeated. This allows us to identify cases where only part of a training example is duplicated (§4.1). *Approximate* full document matching uses hash-based techniques (Broder, 1997) to identify pairs of documents with high n-gram overlap (§4.2).

We identify four distinct advantages to training on datasets that have been thoroughly deduplicated.

Deduplication allows models to emit memorized text; decreases training time; and improves accuracy

Quantifying the Effects of Text Duplication on Semantic Models

Alexandra Schofield[1] **Laure Thompson**[1] **David Mimno**[2]
1 Department of Computer Science, Cornell University, Ithaca, NY
{xanda, laurejt}@cs.cornell.edu
2 Department of Information Science, Cornell University, Ithaca, NY
mimno@cornell.edu

Abstract

Duplicate documents are a pervasive problem in text datasets and can have a strong effect on unsupervised models. Methods to remove duplicate texts are typically heuristic or very expensive, so it is vital to know when and why they are needed. We measure the sensitivity of two latent semantic methods to the presence of different levels of document repetition. By artificially creating different forms of duplicate text we confirm several hypotheses about how repeated text impacts models. While a small amount of duplication is tolerable, substantial over-representation of subsets of the text may overwhelm meaningful topical patterns.

value fit on repeated texts, or even "leak" held out data that is duplicated in the training data. At best, duplication may cause us to overestimate the expressiveness and reliability of models. At worst, models skewed by text duplication may invalidate any conclusions drawn from them, and, by extension, the method itself.

Text replication is a persistent and difficult problem in natural language corpora. In social media settings, partial duplication due to quotation and threading is ubiquitous. Of the 20k posts in the 20 Newsgroups corpus (Lang, 1995), 1151 are exact duplicates, and 25% of the remaining tokens are quoted text from other newsgroup messages.[1] In literary corpora, different versions of the same document may also conflict: text files for Hamlet may differ slightly due to publisher information, line numbers, editorial changes be-

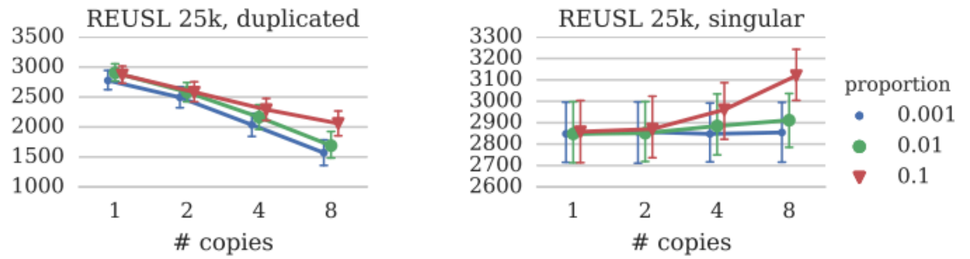


Figure 1: Training perplexity with LDA models trained on the REUSL 25k corpus with 80 topics. Perplexity decreases significantly for the duplicated documents with repetition, but the effect on singular documents is negligible with repeated proportion of the corpus smaller than 0.1.

Why quantify text reuse/duplication?

- Allows us to ask socially relevant questions:
  - Who borrows what from whom and to what end?

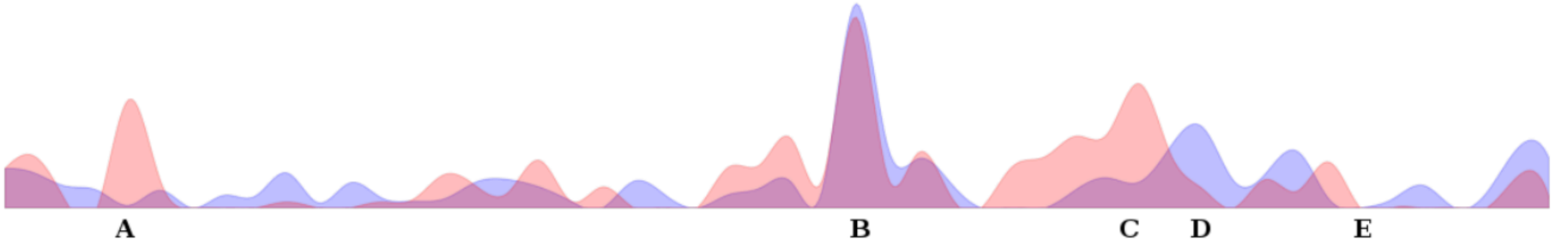


Figure 1: Volume of quotations for each word from a fragment of the 2010 State of the Union Address split by political leaning: conservative outlets shown in red and liberal outlets shown in blue. Quotes from the marked positions are reproduced in Table 1 and shown in the QUOTUS visualization in Figure 2.

Position	Quote from the 2010 State of the Union Address
A	And in the last year, hundreds of al Qaeda's fighters and affiliates, including many senior leaders, have been captured or killed—far more than in 2008.
B	I will work with Congress and our military to finally repeal the law that denies gay Americans the right to serve the country they love because of who they are. It's the right thing to do.
C	Each time lobbyists game the system or politicians tear each other down instead of lifting this country up, we lose faith. The more that TV pundits reduce serious debates to silly arguments, big issues into sound bites, our citizens turn away.
D	Democracy in a nation of 300 million people can be noisy and messy and complicated. And when you try to do big things and make big changes, it stirs passions and controversy. That's just how it is.
E	But I wake up every day knowing that they are nothing compared to the setbacks that families all across this country have faced this year.

Niculae, Vlad, et al. "Quotus: The structure of political media coverage as revealed by quoting patterns." *Proceedings of the 24th International Conference on World Wide Web*. 2015.

Types of text reuse

- Direct Vs. Indirect Quotations
- Paraphrasing
- Summarizing
- Verbatim and Non-Verbatim copying

Our focus

- Many ways to identify duplicated text
- More general problem: quantify the similarity between pairs of texts.
- We will focus on the following:
  - Detect (near) duplicates
  - Sequence alignment

Duplicates

This is a sentence

This is a sentence

If we break a textual sequence into pieces (say words) then duplicates are pairs that have the same pieces and in the same order

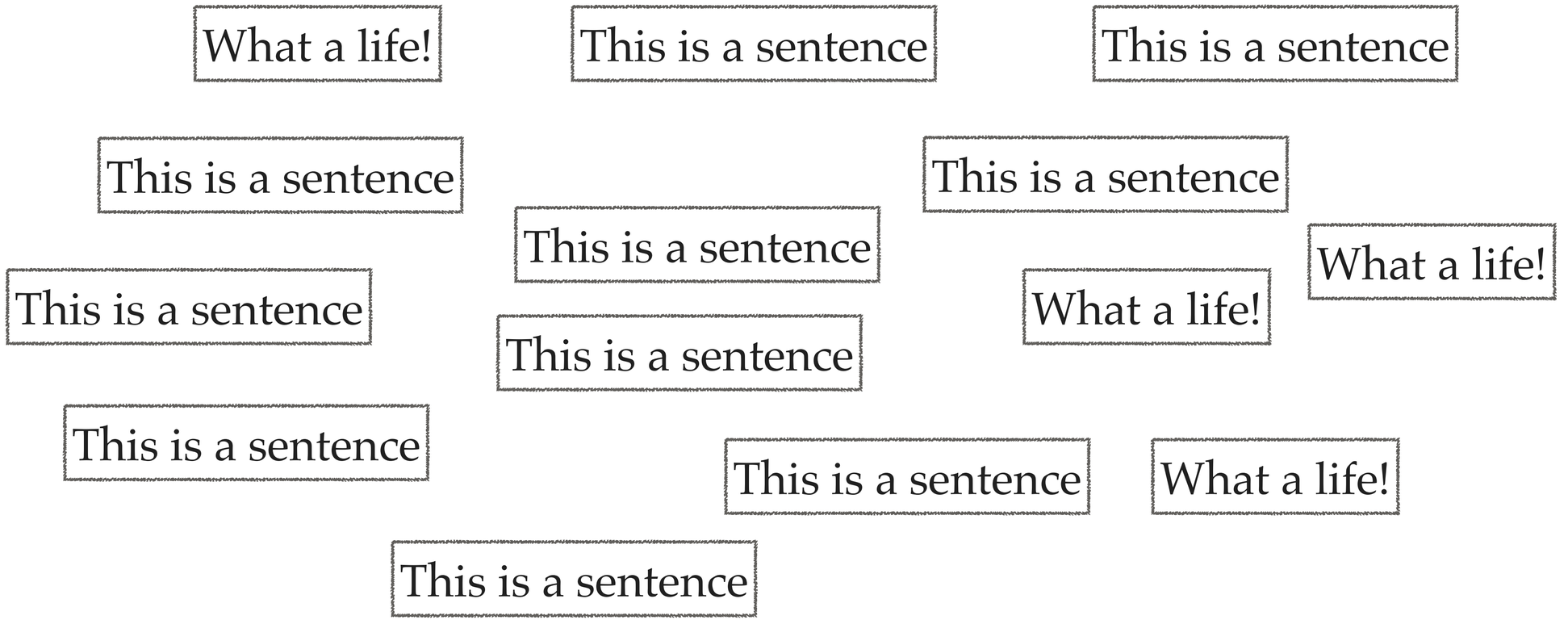
Duplicates

This is a sentence

This is a sentence

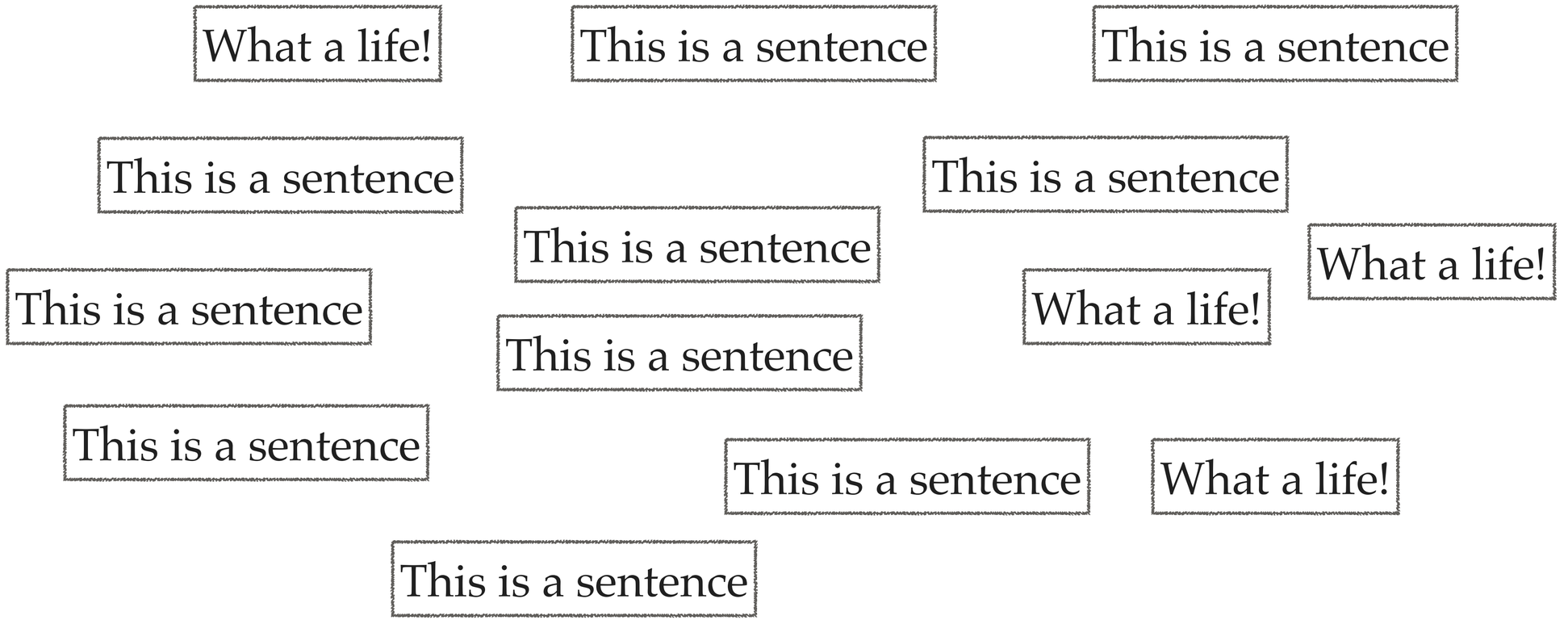
One way to find duplicates in a text corpus is to do pairwise comparisons

Duplicates



In a large corpus of n documents, we end up making many many comparisons

Duplicates



In a large corpus of n documents, we end up making many many comparisons

Duplicates

What a life!

This is a really really really long sentence

This is a really really really long sentence

What a life!

This is a sentence

What a life!

This is a really really really long sentence

This is a really really really long sentence

This is a sentence

What a life!

This is a sentence

The computational cost of pairwise comparison is also dependent on the length of the document

HASHING

- Consider a function h, called a hash function, that takes input as text and produces a hash or a signature
- Same documents produce the same signature, so duplicate documents can be isolated by maintaining a hashtable
- Fast and scales well to document length
- Examples: md5 or SHA, n-gram signatures

Near Duplicates

What a life!

This is a really really really long sentence

This is a really really really long sentence

What a life!

This is a sentence

What a life!

This is a really really really long sentence

This is a really really really long sentence

This is a good sentence

What a life!

This is a good sentence

Ideally we want to group duplicates and almost duplicates together

Locality sensitive hashing

- Design a hash function h:
  - h maps similar documents into the same bucket with high probability
  - h maps dissimilar documents into the same bucket with low probability
- There are many ways to come up with such hash functions

Locality sensitive hashing

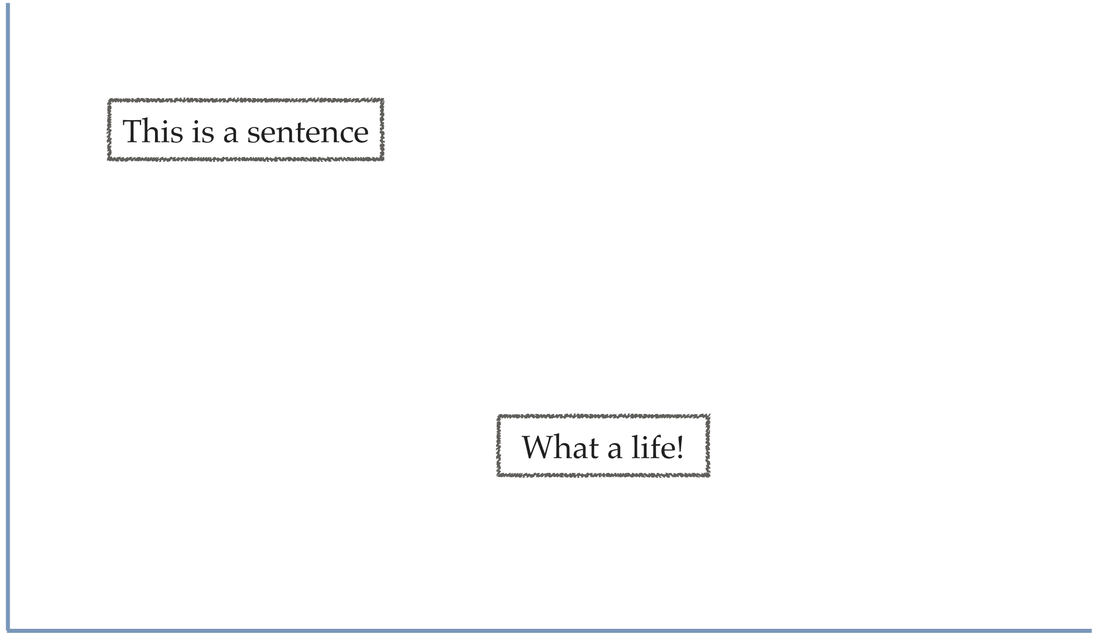
This is a sentence

What a life!

Locality sensitive hashing

This is a sentence

What a life!



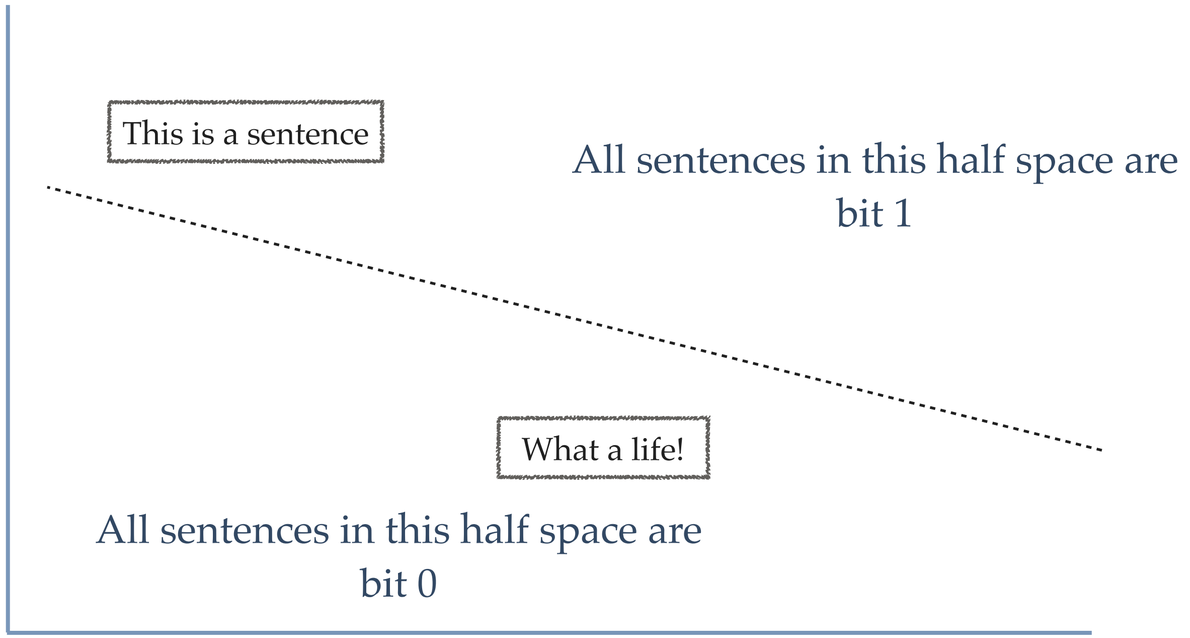
Locality sensitive hashing

This is a sentence

1

What a life!

0



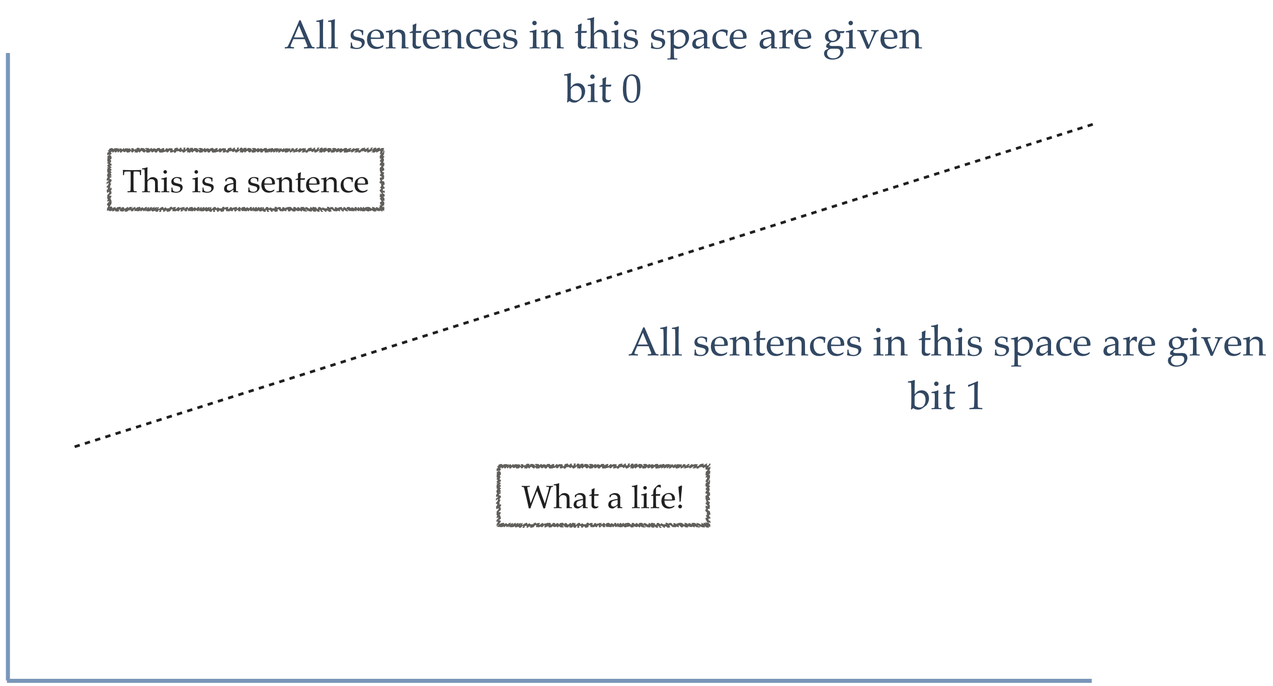
Locality sensitive hashing

This is a sentence

1 0

What a life!

0 1



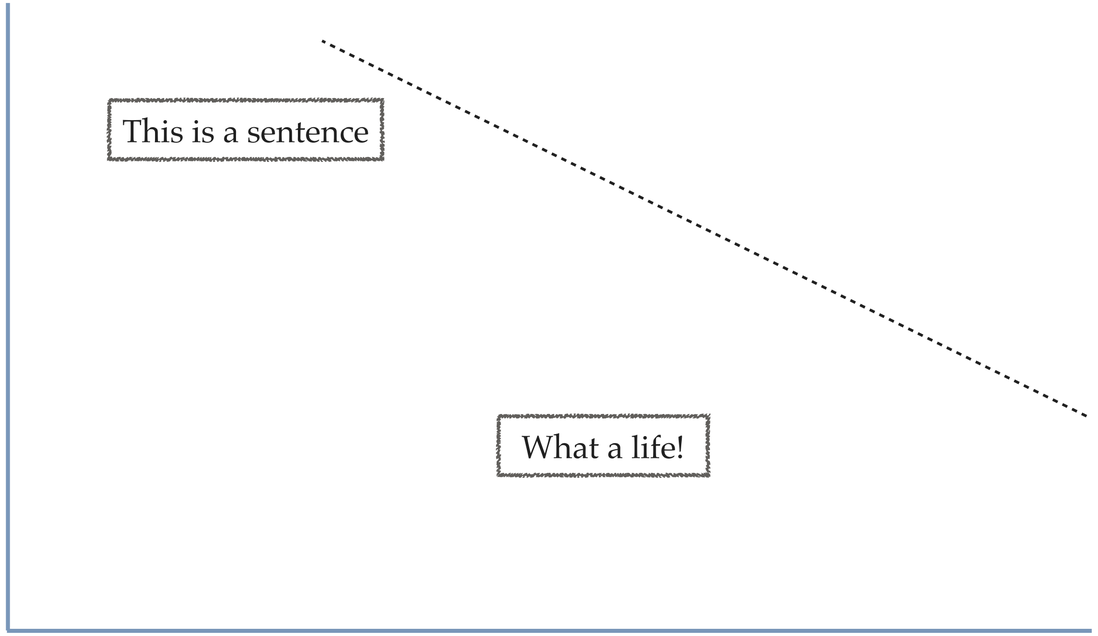
Locality sensitive hashing

This is a sentence

1 0 0

What a life!

0 1 0



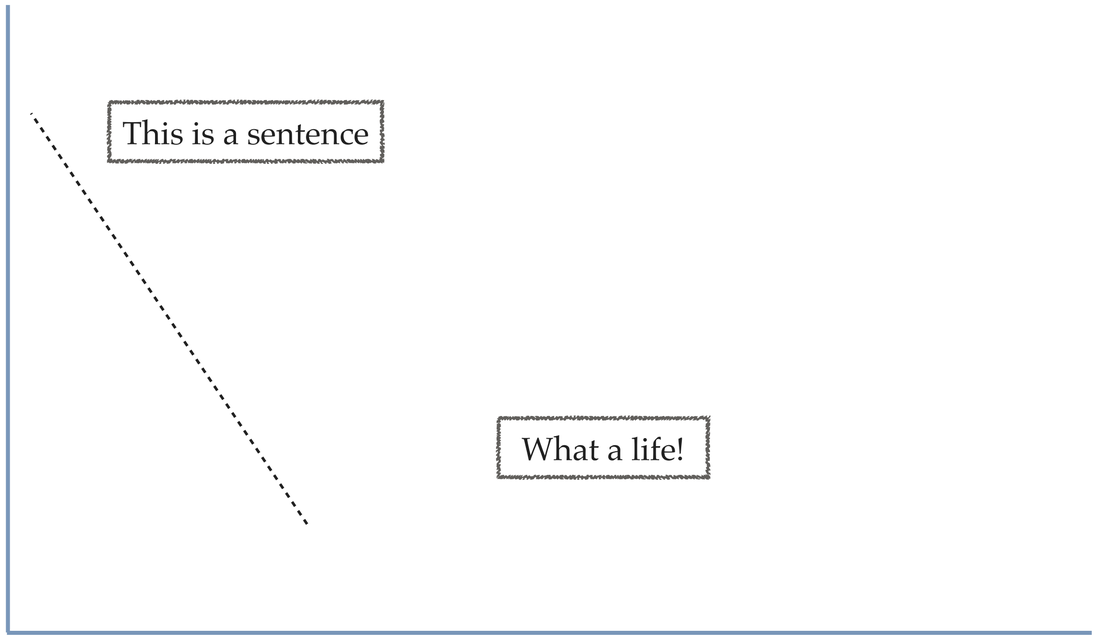
Locality sensitive hashing

This is a sentence

1 0 0 1

What a life!

0 1 0 1



Locality sensitive hashing

This is a sentence

1 0 0 1…0

What a life!

0 1 0 1…0

Locality sensitive hashing

This is a sentence

1 0 0 1…0

What a life!

0 1 0 1…0

It can be proven that this method produces signatures of documents such as similar documents will have similar signatures with high probability

Locality sensitive hashing

- Locality sensitive hashing is useful to find similar documents, to cluster documents, etc

Sequence alignment

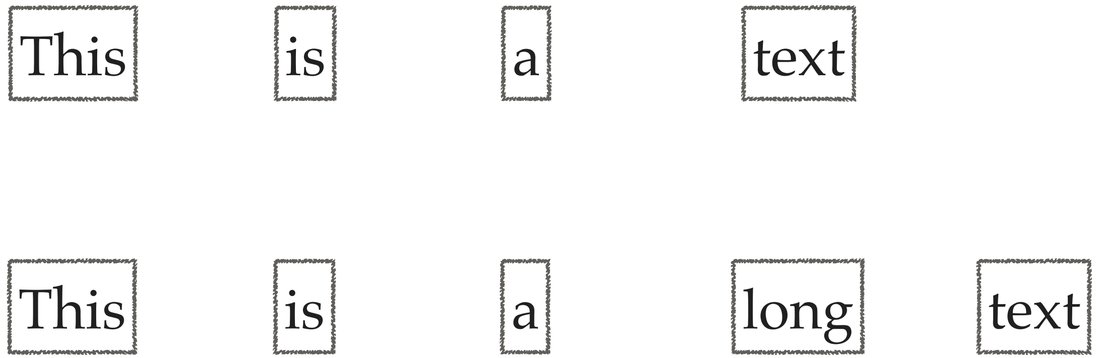
- Another systematic way of quantifying duplication across texts is to assume is to do pairwise comparison

Sequence alignment

This is a text → Noisy channel → This is a long text

Sequences become duplicates or near duplicates after passing through a noisy channel which garbles part of the text

Sequence alignment



Sequence alignment

This	is	a	-	text
This	is	a	long	text

We try to align the sequences such that they maximize the matches

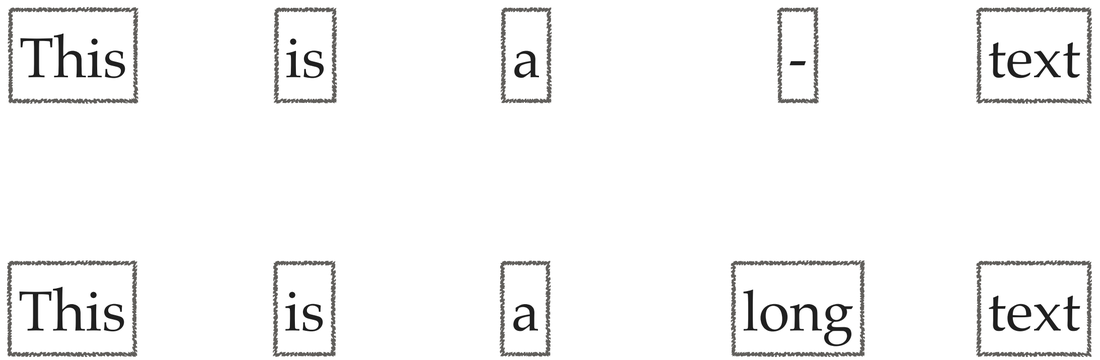
Scoring system

This	is	a	-	text
This	is	a	long	text

Word matches = +1

Word mismatches = -1

Scoring system



Score=3

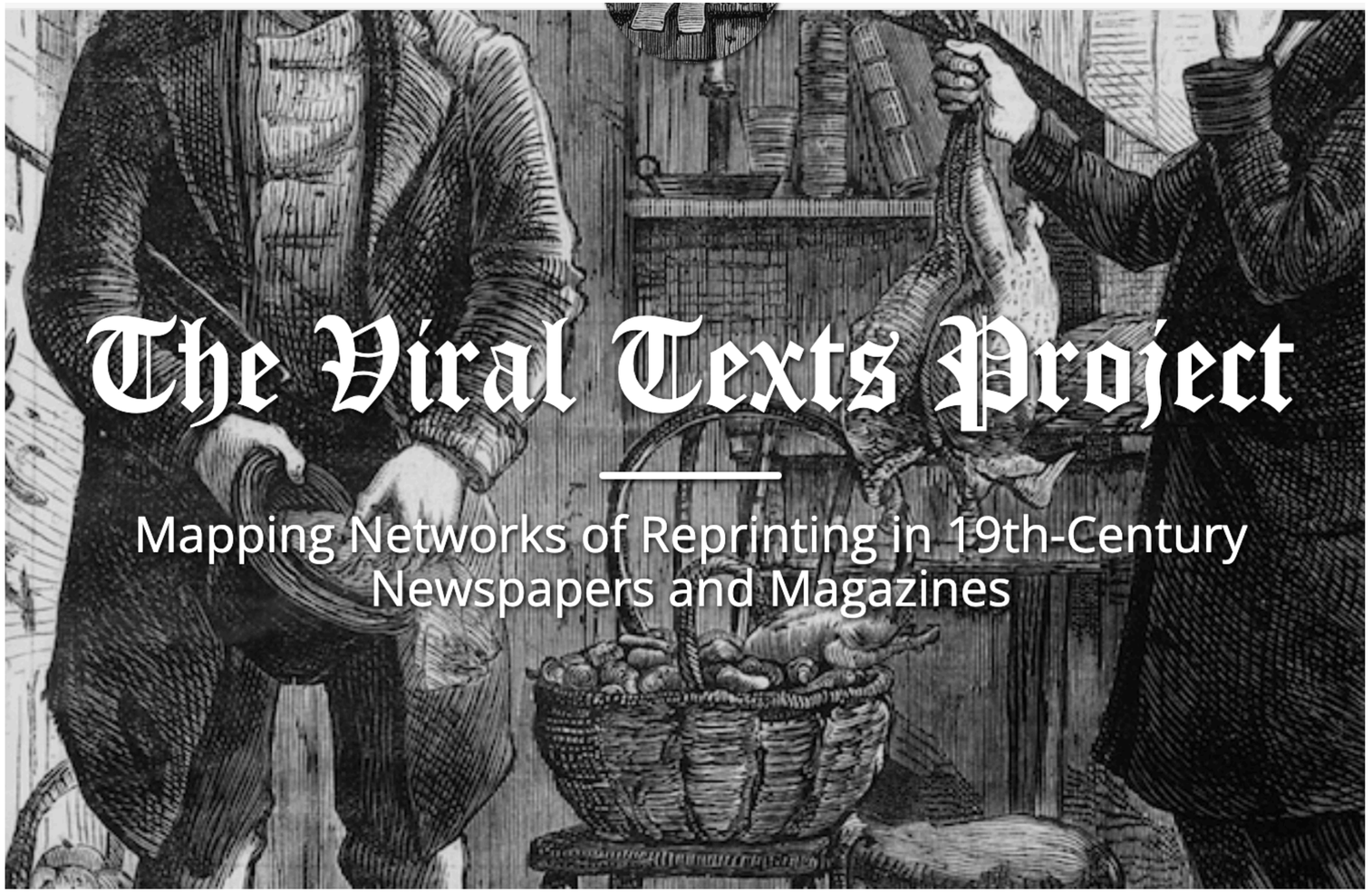
Word matches = +1

Word mismatches = -1

Sequence alignment

- Many ways to quantify distances between sequences
- Algorithms usually compute the distance using dynamic programming
- Variations include aligning pairs or aligning multiple sequences at the same time
- Example: Smith-Waterman, Needleman-Wunsch

What can text reuse tell us?



> *"Nothing but a newspaper can drop the same thought into a thousand minds at the same moment..."*
>
> *—Alexis de Tocqueville, Democracy in America*

The Raftsman's Journa

BY S. J. ROW. CLEARFIELD, PA., WEDNESDAY, NOVEMBER 4, 1868. VOL

Select Poetry.

MARJORIE'S ALMANAC.

Robins in the tree tops,
Blossoms in the grass;
Green things a-growing
Everywhere you pass;
Sudden little breezes,
Showers of silver dew,
Black bough and bent twig
Budding out anew;
Pine tree and willow tree,
Fringed elm and larch—
Don't you think that May-time's
Pleasanter than March?

Apples in the orchard,
Mellowing one by one;
Strawberries upturning
Soft cheeks to the sun;
Roses faint with sweetness,
Lillies, fair of face;
Drowsy scents and murmurs
Haunting every place;
Lengths of golden sunshine,
Moonlight bright as day—
Don't you think that summer's
Pleasanter than May?

Roger in the corn patch
Whistling negro songs;
Pussy by the hearth side,
Romping with the tongs;
Chestnuts in the ashes,
Bursting through the rind;
Red leaf and gold leaf
Rustling down the wind;
Mother "doin' peaches"
All the afternoon—
Don't you think that Autumn's
Pleasanter than June?

Little fairy snow flakes
Dancing in the flue,
Old Mr. Santa Claus,
What is keeping you?
Twilight and fire-light,
Shadows come and go;
Merry chime of sleigh bells,
Tinkling through the snow;
Mother's knitting stockings,
(Pussy's got the ball)—
Don't you think that Winter's
Pleasanter than all?

THE DASHFORD TRAGEDY.

Dashford had a sensation, and it was so new a thing that all the village was agog—with cars to hear of, and eyes to view the object. Everybody in the place, probably, who could read, had read the name on the books of the Red Mug—Charles Wylie, New York City.

The Red Mug was the centre of attraction. After his sign was put out it was positively alarming to observe how unhealthy Dashford suddenly became. Hitherto people, for the most part, had died either by accidents or from old age—but now the entire female community had gone ill. Coughs, colds, nervous diseases, fevers, and disordered livers was the rule, and not the exception.

Dr. Wylie was kept riding for the greater part of the time, and the principal wonder was when the poor fellow contrived to obtain any sleep. He was an immense favorite with the ladies, both old and young. He had such sad eyes when his countenance was at rest that they were sure he must have some secret trouble—and there is no surer method for a man to make himself interesting than to give people the impression that he is bearing in silence some great sorrow.

Though polite and courteous to all, Dr. Wylie was not long in making his selection, and it did infinite credit to his good taste.

Lucy Walbridge was by far the sweetest girl in Dashford. She was about twenty-five years of age—an orphan and an heiress, and resided with her uncle, Squire Hillman, at the Hall. And Squire Hillman's wife was obligingly taken sick of a slow fever, which gave the Doctor an excellent excuse for tying his roan horse, every day, to the great elm in front of the Squire's.

We are not writing a love story, so we will pass over the courtship. For once the course of true love seemed to run smooth. There were no obstacles to surmount—both parties were of an age to marry—and there were no friends to raise objections.

It was in July that Dr. Wylie came to Dashford, and his wedding day was set for the 15th of March.

It came all too soon, Lucy thought, for surely nothing could be more delightful than the charmed life they were leading. She almost feared marriage might break the sweet enchantment.

The day was clear and cloudless, altogether unlike the days March usually gives us, and in the morning the first bluebird sang gaily in the old elm, which reached its branches almost in at Lucy's window. Dr. Wylie made all his business calls—for the sick must be attended to—and on his way to his hotel, he stopped at the Hall, in defiance of all etiquette, to kiss Lucy and bid her keep up her courage. He ate his supper with Mrs. Stark at six—and then went to his room to dress. The ceremony was to take

tives in New York, who were at once written to.

As is usual in such cases, public indignation ran very high. Every one was anxious to convict the real assassin, that the vengeance might be swift and sure. Dr. Wylie's brother offered a reward of five hundred dollars for the discovery and apprehension of the murderer; and Dashford, not to be behind in the good work, offered a like amount.

The offered rewards brought forth their fruit. Isaac Smith, a laborer, employed at intervals about the Red Mug, came before a justice and stated that on the evening of the murder, about six o'clock, he had met Clyde Irving—a young mechanic—coming in great haste from the direction of the garden at the Red Mug. He had bidden him good evening, a salutation which was briefly responded to. Irving had appeared to be powerfully agitated from some cause, and anxious to escape. The next morning, feeling curious, with the rest, about the murder and everything connected with it, Mr. Smith had been over the garden, and on looking beneath the hemlock which covered the sage bed, he had found a small, exceeding sharp chisel, bearing on the handle the name of Clyde Irving. The instrument was rusty and stained with blood as he exhibited it to the justice—and the finding of this weapon recalled the fact that, at the *post mortem* examination, the surgeon had expressed it as his opinion that the fatal wound had not been made by a knife, but by some other sharp pointed instrument.

Clyde Irving was a young man of irreproachable habits—poor, but honest, temperate and virtuous—and the only son of a widowed mother. It was well known in Dashford that he had loved Lucy Walbridge from childhood, but that he had not spoken to her on the subject, because his station in life was so much below hers.

You all know how readily people find reasons for the truth of what they desire to believe. Irving had not an enemy in the village—but still it was necessary to have some one on whom to throw the guilt, and they were all glad that the murderer had been discovered. A score of trifling circumstances were brought against the unfortunate young man, and he was arrested, tried and convicted of the murder of Charles Wylie, on the evening of the 15th of March.

Lucy, who had in a measure recovered

ly stream. Over the dead body which they brought home to me I swore an oath—that before Charles Wylie should marry any woman he should taste death! I have kept the oath. With this hand I murdered him—striking the fatal blow with a chisel I obtained at Clyde Irving's shop, where I called to make some trifling inquiry. I deserved death! I think God, who knows every tried and tempted heart, will judge me leniently. Oh, my soul shudders when I remember the hearts he has desolated—the hearths he has laid waste—for my Alice was only one of many victims!

"I killed him and escaped through the window. In leaving the garden I saw Clyde Irving there—I think, for some reason, he had a distrust of me; but as there was nothing to confirm it he kept it to himself."

She paused, but though all present believed her story, not a man of them lifted a hand to deprive her of freedom.

The Sheriff unbound Clyde, and allowed him to descend the scaffold. He was free.

At last one of the constables approached Mrs. Sinclair, who, with bowed face, was leaning against the railing of the scaffold. She lifted her head, divining his purpose, and waved him back. "The law has no power over the dead," she said hoarsely; "I am free!"

Even as she spoke her lips grew purple—she tottered and fell forward; and before they reached her she was lifeless. An examination after death proved that she had swallowed strychine—and they buried her and her sins together in the village churchyard.

Two years afterward, Clyde Irving married Lucy Walbridge.

KEEP WARM AND SAVE YOUR LIFE.—At this season many deaths take place which might be prevented by warmer clothing. Many a fatal case of dysentery is caused by the want of a woolen undershirt, or of an extra blanket at night. The sudden changes of the tempeasture which occur at this period of the year are very trying to the constitution. People with weak lungs quickly feel the effect of them. Frequently the thermometer falls many degrees within a few hours. Not only the feeble, but robust and strong persons suffer from such great variations of temperature. When the weather grows cold rapidly the pores of

A "stunning" Love Letter.

The following is sublimely 'splendiferous,' and we recommend it as a model to letter writers:

MY DEAR MISS C.—Every time I think of you my heart flops up and down like a churn dasher. Sensations of unutterable joy creep over it like young goats over a stable roof, and thrill through it like Spanish needles through a pair of tow linen trowsers. As a gosling swimeth with delight in a mud-puddle, so swim I in a sea of glory. Visions of ecstatic rapture, thicker than the hairs in a blacking brush, and brighter than the hue of the humming bird's pinions, visit me in my slumbers, and borne on their visible wings, your image stands before me, and I reach out to grasp it, like a *pointer* snapping at a blue bottle fly. When I first beheld your angelic perfections, I was bewildered, and my brain whirled about like a bumble bee under a glass tumbler. My eyes stood open like cellar doors in a country town, and I lifted up my ears to catch the silvery accents of your voice. My tongue refused to wag, and in silent adoration I drank in the sweet infection of love as a thirsty man swalloweth a hot whiskey punch. Since the light of your face fell upon my life, I sometimes feel as if I could lift myself up by my boot straps to the top of the presbyterian steeple, and pull the bell rope for singing school. Day and night you are in my thoughts. When Aurora, blushing like a bride, rises from her saffron couch; and when the jay bird pipes his tuneful lay in the apple tree, by the spring house; when the chanticleer's shrill clarion heralds the coming morn; when the awakened pig ariseth from his bed and grunteth, and goeth for his morning refreshments; when the drowsy beetle wheels his drowning flight, at sultry noontide, and when the lowing cows come home at milking time, I think of thee; and like a piece of gum elastic, my heart seems to stretch clean across my bosom. Your hair is like the mane of a sorrel horse, powdered with gold; and the brass pin skewered through your waterfall, fills me with unbounded awe. Your forehead is smoothar than the elbow of an old coat. Your eyes are glorious to behold. In their liquid depths I see legions of little Cupids, bathing like a cohort of ants in an old army cracker. When their fire hit upon my manly breast, it penetrated my entire anat-

Who Ate Roger Williams?

We take the following from *Steele's Fourteen Weeks in Chemistry*: The truth that matter passes from the animal back to the vegetable, and from the vegetable to the animal kingdom again, received a curious illustration not long since.

For the purpose of erecting a suitable monument in memory of Roger Williams, the founder of Rhode Island, his private burying ground was searched for the graves of himself and wife. It was found that everything had passed into oblivion. The shape of the coffins could only be traced by a black line of carbonaceous matter. The rusting hinges and nails, and a round wooden knot, alone remained in one grave, while a lock of braided hair was found in the other. Near the grave stood an apple tree. This had sent down two main roots into the very presence of the confined dead. The larger root, pushing its way to the precise spot occupied by the skull of Roger Williams, had made a turn as if passing around it, and followed the direction of the back bone to the hips. Here it divided into two branches, sending one along each leg to the heels, when both turned upward to the toes. One of these roots formed a slight crook at the knee, which made the whole bear a striking resemblance to the human form. These were the graves, but their occupants had disappeared; the bones had even vanished. There stood the thief—the guilty apple tree—caught in the very act of robbery. The spoliation was complete. The organic matter, the flesh, the bones of Roger Williams had passed into an apple tree. The elements had been absorbed by the roots, transmuted into woody fiber, which could now be burned as fuel, or carved into ornaments, had bloomed into fragrant blossoms, which delighted the eye of the passer-by, and scattered the sweetest perfume of spring; more than that—has been converted into luscious fruit, which from year to year had been gathered and eaten. How pertinent then, is the question, "Who ate Roger Williams?"

FUNNY SCENE IN COURT.—The Judge of one of the New Orleans municipal courts sat gloomy and grand on his bench of ermine. The prisoner occupied the dock, apparently meek and downcast. She had a merry twinkle in her eye, however, that

HOOFLAN

HOOFL

For all diseases

Hoofland

Is composed
medicinally te
Barks, making
trated, and e
mixture of any

HOOFLA

Is a combinatio
tern, with the p
Orange, &c., m
agreeable reme

Those prefer
ic admixture, w

HOOFLAN

Those who ha
of the Bitters, a

HOOFLA

They are bot
same medicinal
two being a mer
the most palatal

The stomach,
Indigestion, Dy
very apt to ha
Liver, sympathi
the Stomach, the
of which is that
or more of the f

Constipation, Fl
of Blood to th
Nausea, Heart
or Weight in
Sinking or Flu
Swimming of
Breathing, Flu
Suffocating Se
Dimness of Vis
Dull Pain in t
tion, Yellowne
the Side, Back
es of Heat, Bu
agginings of Ev

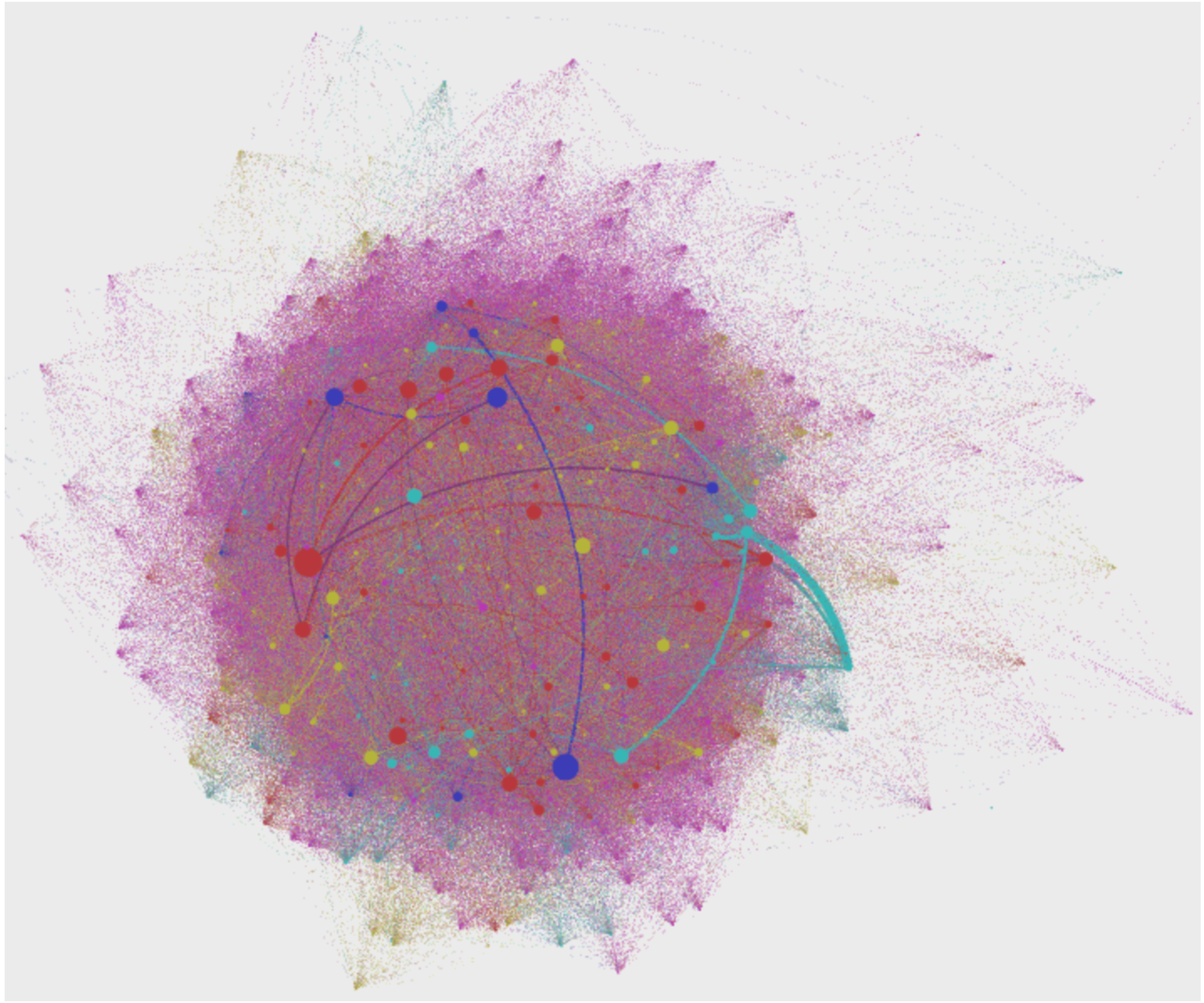
The sufferer fr
the greatest caut
for his case, pure
sured from his i
possesses true m
ed, is free from
established for i
these diseases.
submit those wh

Hoofland's Ge
German To
Jacks

Twenty-two ye
duced into this
which time they
more cures, and
a greater extent
to the public.

Texts of all kind
(e.g., news, advertisements, jokes, stories)
were shared across newspapers

https://loveletter.viraltexts.org/neatline/fullscreen/exhibit



Allows the construction of a newspaper network and identify items that became viral in the 19th century

Popularity

- Engagement metrics on many social media platforms indicate the popularity of the content

Impact

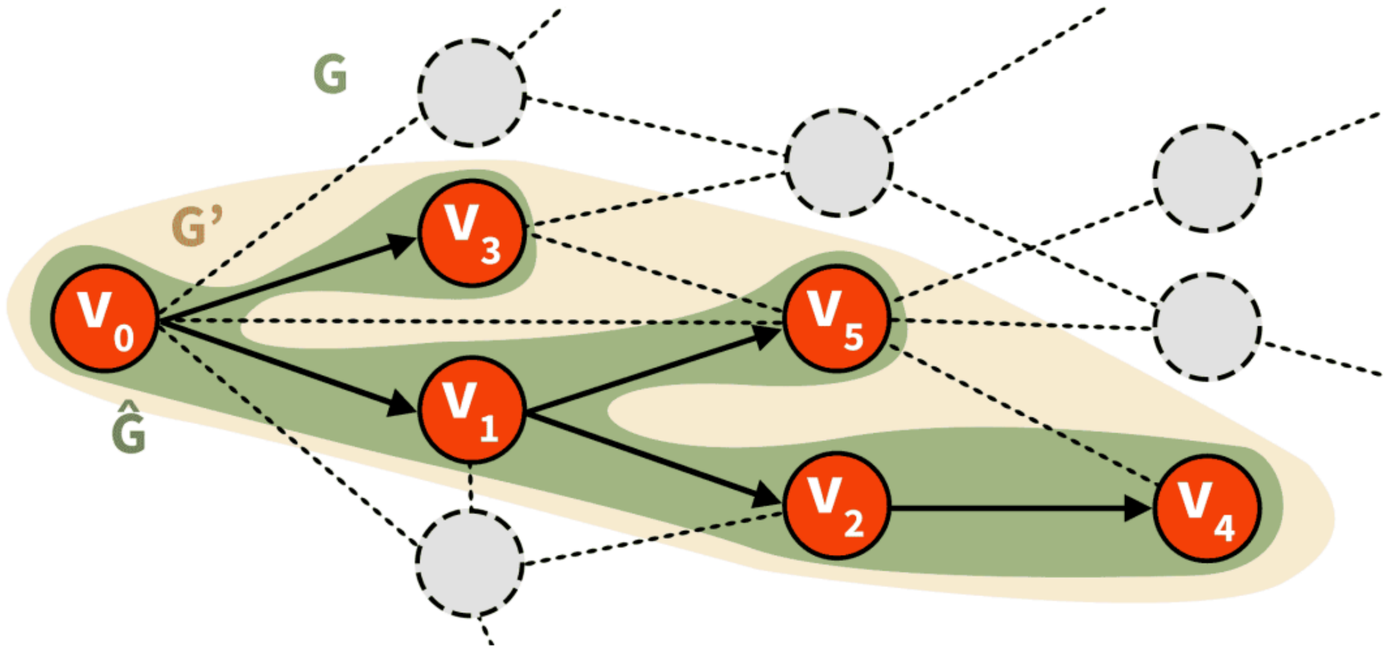
- In some domains, the content is referred, indicating the importance of the content

Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… We introduce **BERT** and its detailed implementation in this … For finetuning, the **BERT** model is first initialized with the pre-… A distinctive feature of **BERT** is its unified architecture across …
★ Save Cite Cited by 83562 Related articles All 46 versions »

“What factors make text popular/viral?”

Factors affecting popularity

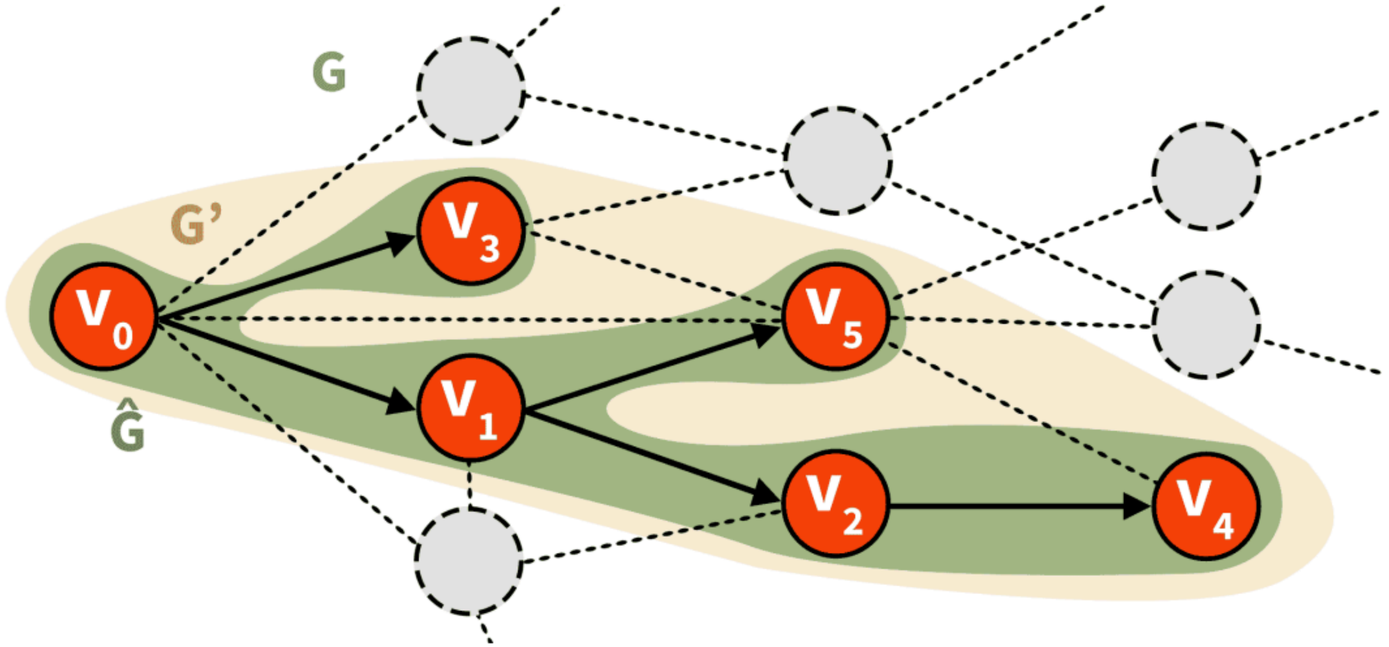
- Content



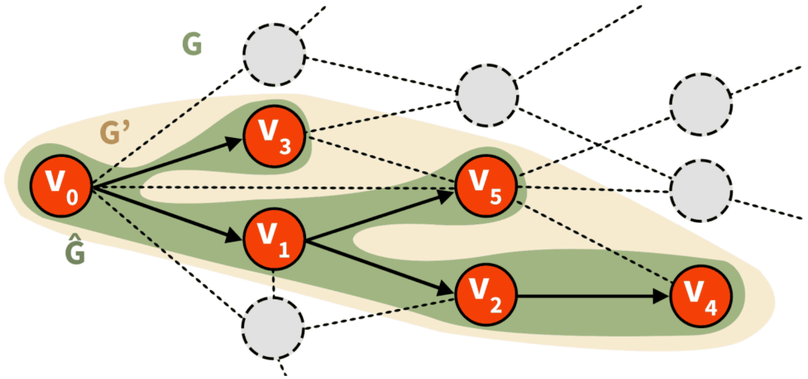
If we observe a Facebook post reshared k times, can we predict if the cascade will reach a certain size?

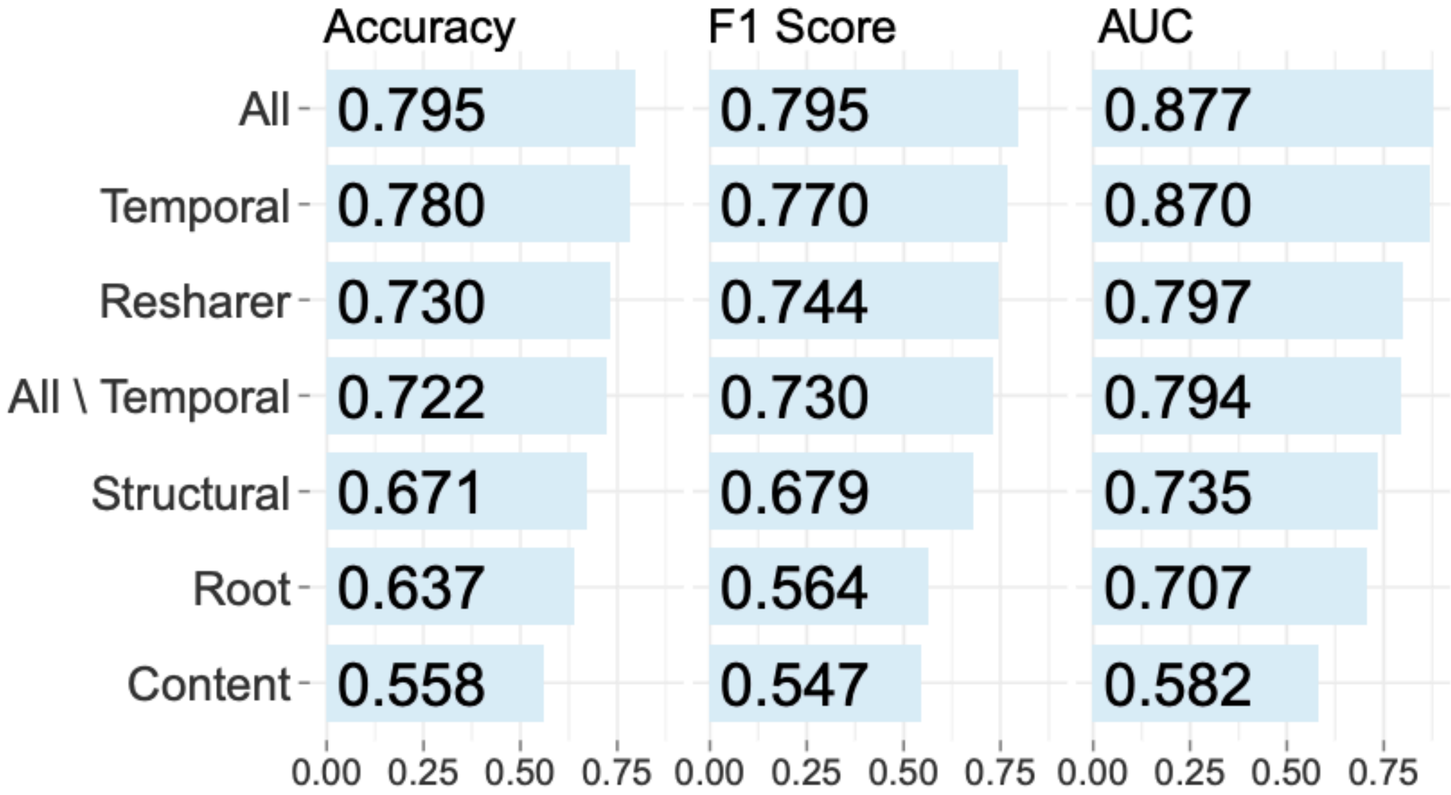
Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. 2014.

A cascade is a temporal sequence of events such as resharing of a post



Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. 2014.





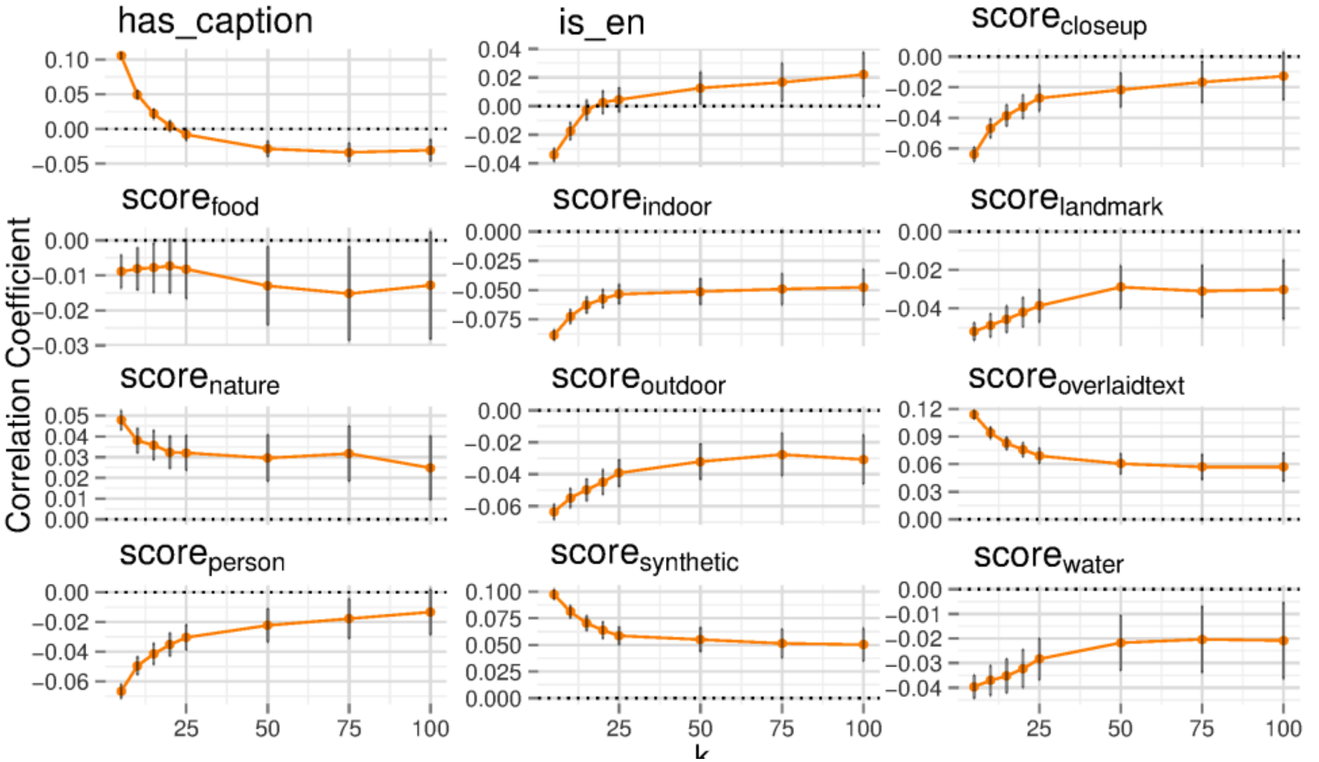
Content is an important factor in a post being reshared

Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. 2014.

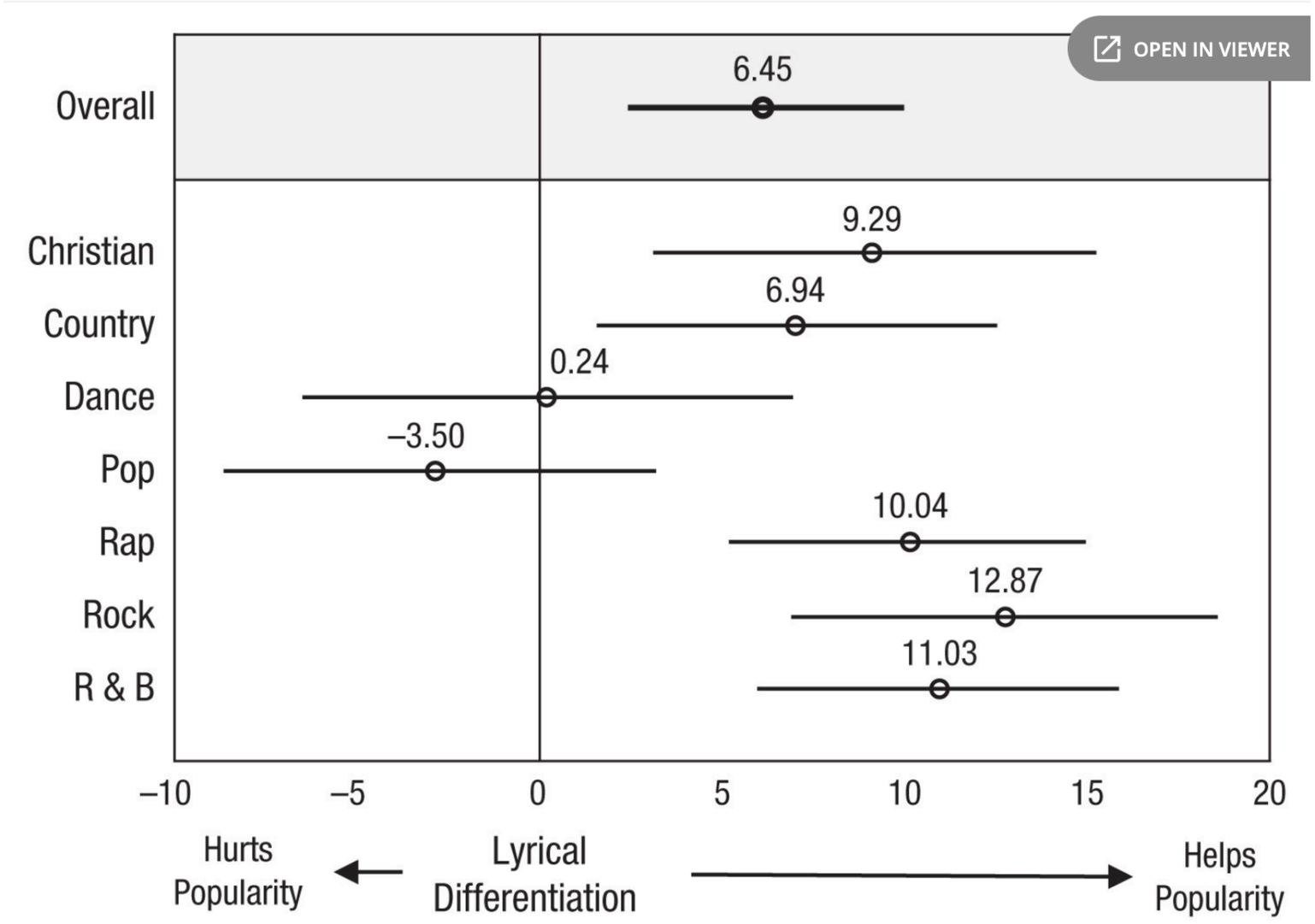
Content Features

$score_{food/nature/...}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
is_en	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$liwc_{pos/neg/soc}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English

(a) Content



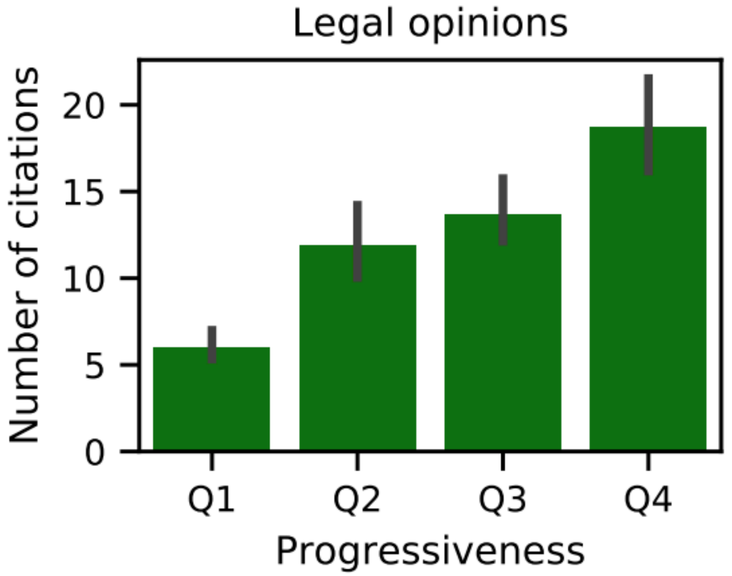
Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. 2014.

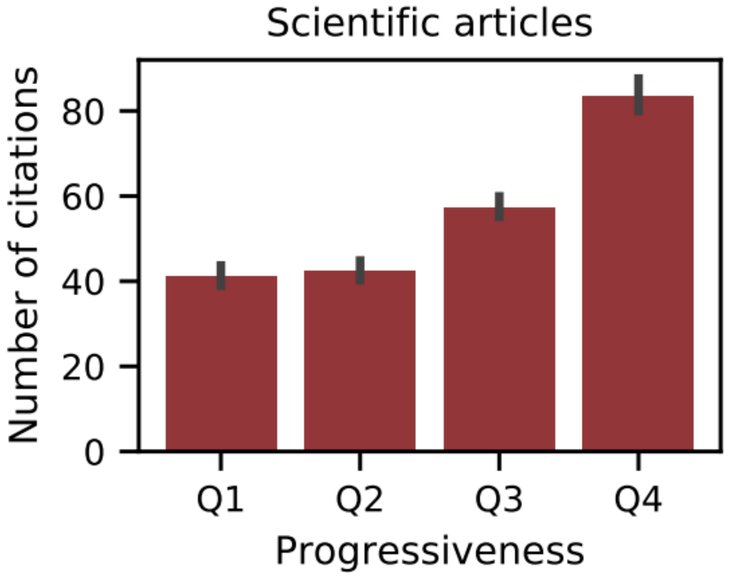


What features of a song lyric makes it popular?

Songs that are atypical of their genre tend to be more popular

Berger, Jonah, and Grant Packard. "Are atypical things more popular?." *Psychological science* 29.7 (2018): 1178-1184.





Which papers/opinions have more impact?

Papers/opinions that are "ahead of their time"

Soni, Sandeep, Kristina Lerman, and Jacob Eisenstein. "Follow the leader: Documents on the leading edge of semantic change get more citations." *Journal of the Association for Information Science and Technology* 72.4 (2021): 478-492

Factors affecting popularity

- Social connections

Factors affecting popularity

- Social connections
  - We are generally part of a social network, so information flows from person to person
  - Position in a network and influence can determine popularity

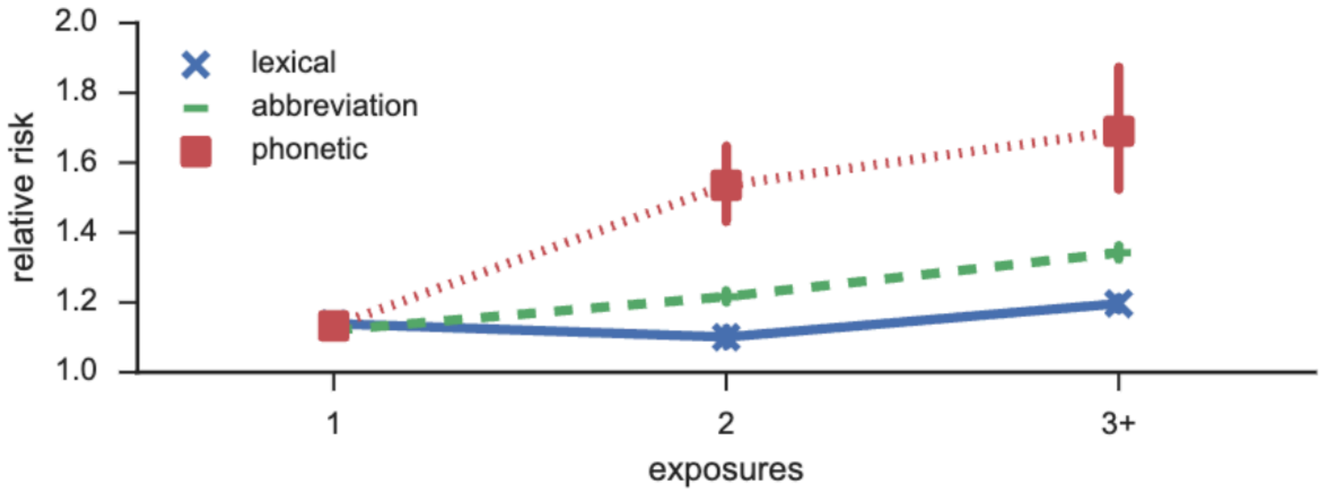
Epidemic spread

- Many diseases can be modeled as person-person interactions
  - Exposure: A person comes in contact with another person who carries a virus
  - Infection: The person becomes infected due to exposure

Can text popularity be thought of as a contagion?

Language change as Contagion

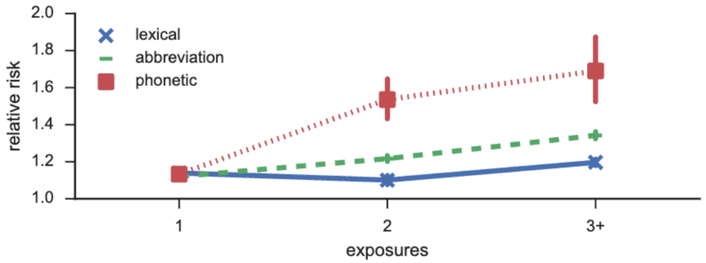
- How many exposures are needed before someone adopts a new word?



Source: Goel et. al. (2016)

Language change as Contagion

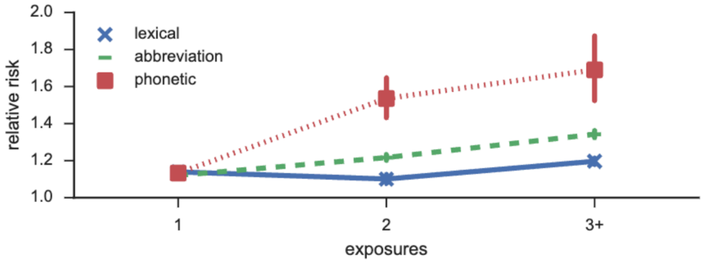
- Only one exposure needed for adoption
  => simple contagion
- Multiple exposures needed for adoption
  => complex contagion



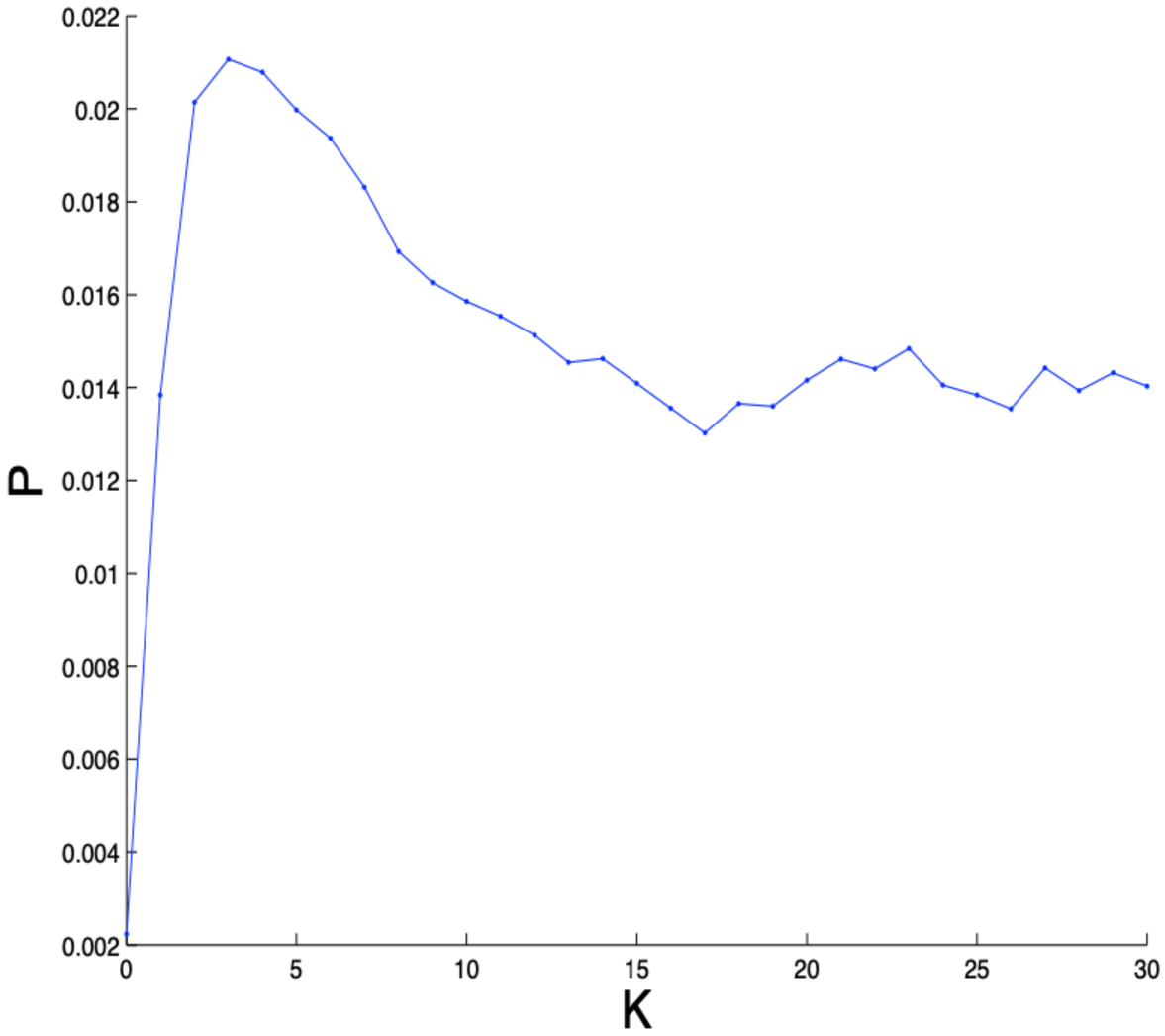
Goel et. al.

Language change as Contagion

- Some words that are more familiar (e.g., hella) spread as a simple contagion
- Others spread that are new (e.g., jawn) require more exposures and hence follow complex contagion



Goel et. al.



Romero et. al. 2011

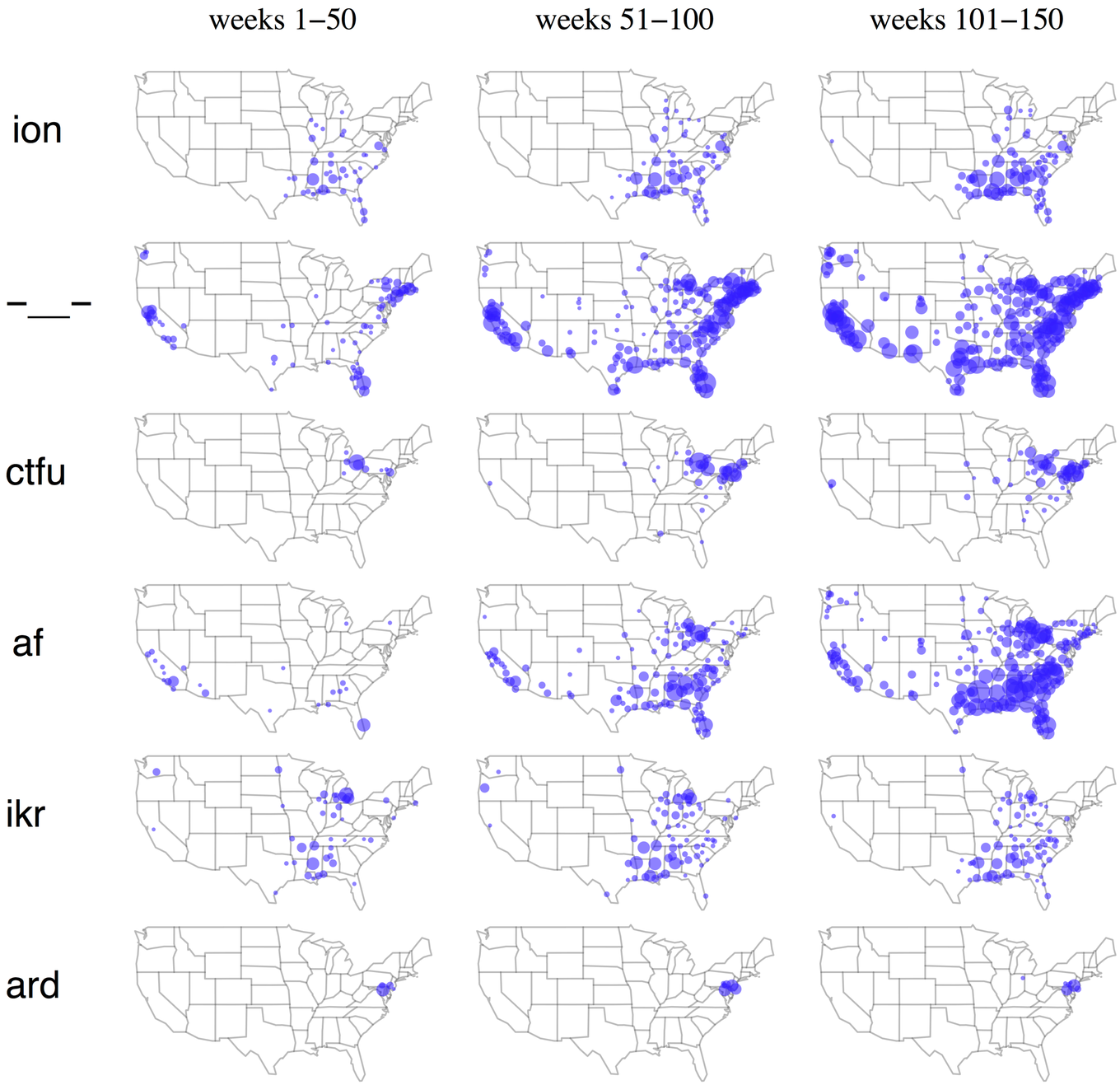
- Political hashtags on Twitter (e.g., #tcot) follow a complex contagion

Factors affecting popularity

- Identity

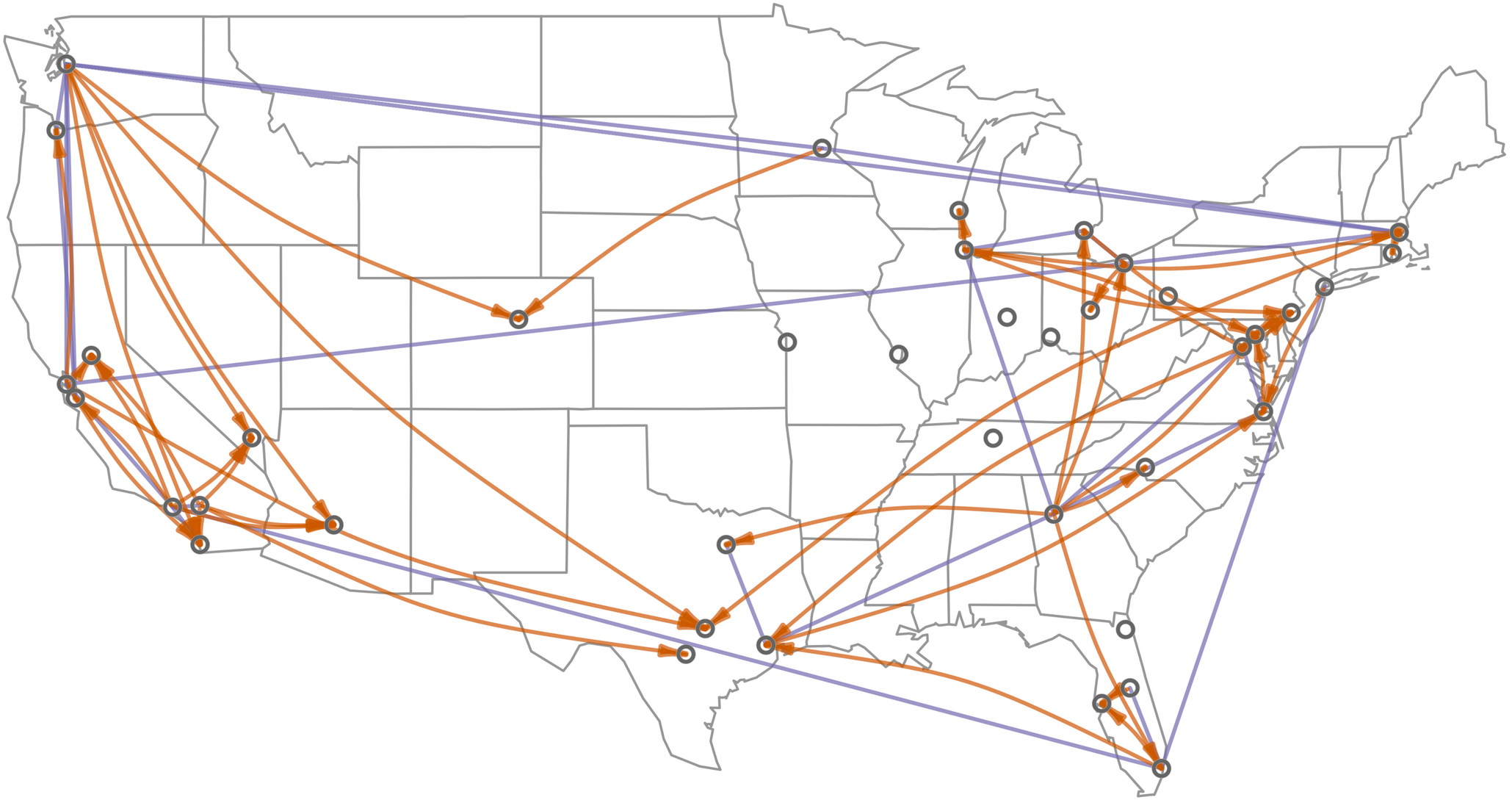
Factors affecting popularity

- Identity
  - Adoption of items is often related to whether the item relates to your own identity and in response to dissimilar others.



Linguistic innovations (e.g., new words) show different diffusion patterns based on

identifying characteristics of users such as their race, age, gender, and geography



Eisenstein et. al.

In class

- String2String: https://github.com/stanfordnlp/string2string
- String2string demo