Søketek Uke 2

Gruppe 1

Agenda

Første time

- Science fair
- Repetisjon
- Mer om Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing

Ideer til science fair

Group	Topic
1	Google's search ranking and how it is affected by their business model
2	Prime factorization and information retrieval
3	Suffix arrays and LCP
4	String search
5	Transformer memory as a differentiable search index
6	HyperLogLog
7	The Rocchio algorithm for query expansion
8	Neural embeddings: The backbone of modern search engines
9	Compression: DEFLATE and Huffman
10	ANDNOT
11	How to minimize confirmation bias when searching
12	Bloom filters
13	Approximate nearest neighbours
14	Support vector machines
15	How to squeeze a lexicon
16	Text classification and MonkeyLearn
17	Strategic considerations for the design of search systems in multi-sided platforms
18	SEO and how to win the battle of the search results
19	Approximate string matching
20	Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs
~~21~~	
22	The Aho-Corasick algorithm
23	Applications of IN4120 in legal

	C	D
	Topic	**Date**
4	MapReduce	
6	Non-Latin alphabets	
2	Alternative language classifiers	
1	Caching	
2	Alternative language classifiers	
3	Skip lists	
5		
5	Tries for approximate string matching	
4	MapReduce	
7	Fuzzy search	
2	Alternative language classifiers	
6	Non-Latin alphabets	
1	Caching	
7	Fuzzy search	
3	Skip lists	

Agenda

Første time

- Science fair
- Repetisjon
- Mer om Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing

Repetisjon

- Parsing
  - Tokenization
  - Normalization og case folding
  - Lemmatization
  - Stemming
- Skip pointers

Tokenization

- Dele opp tekstsekvens til tokens
  - «En gul bil» → «En», «gul», «bil»
- Mange måter å lage tokens
  - «Bilens farge er gul» → Bilen/Bilens/Bilen sin
  - Dette skal ikke dere ta hensyn til

Normalization

- Mappe text og query term til samme format
  - «U.S.A» og «USA»
  - «én» → «en»

Case folding

- Gjøre alle bokstaver lower-case
  - «Kongen av Norge» → «kongen av norge»

Lemmatization

- Redusere varianter av ord til basisformen
  - am, are, is → be
  - car, cars, car’s, cars’ → car

Stemming

- Vi ønsker at forskjellige former av det samme ordet skal matche
- Finne «stammen»: reduce a word to its stem
  - Automates, automatic → automat
- Trenger ikke bli et «ekte» ord
- Porter’s algoritme

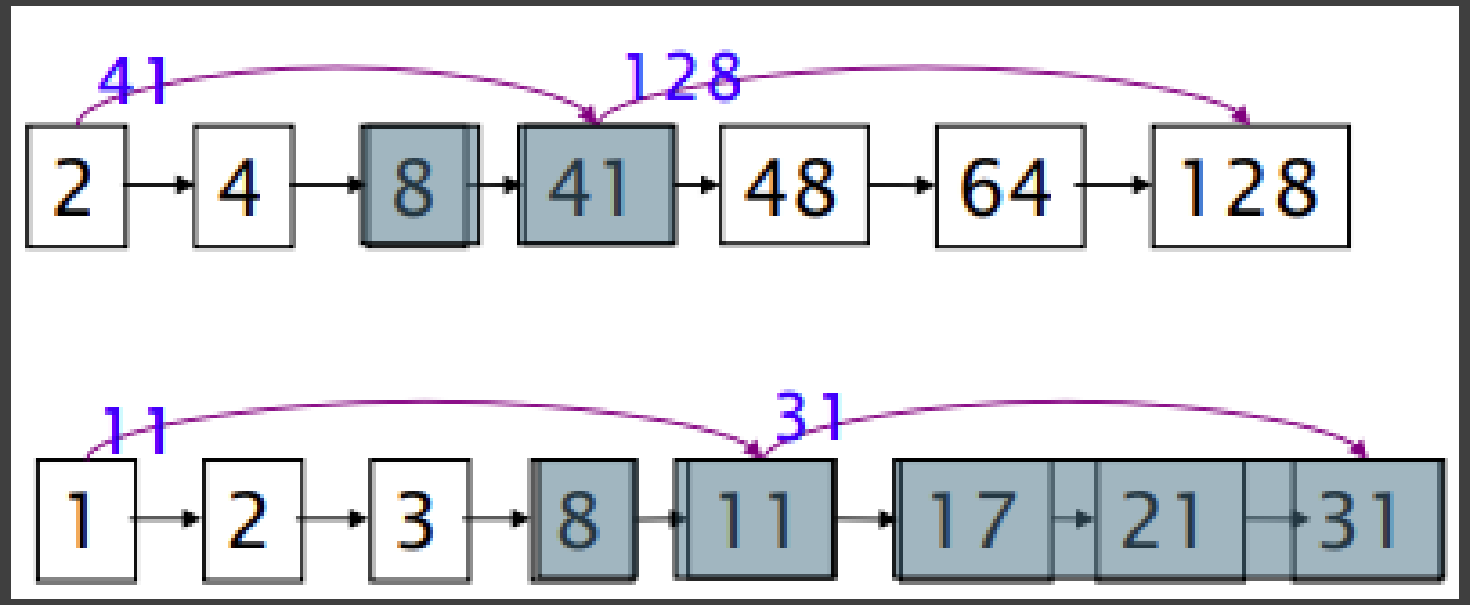
Skip pointers

- Trenger ikke alltid gå gjennom alle postinger
- Unødvendig å sjekke informatikk 58 ganger
- Kan holde referanser til senere postings

```
inverted_index = {
    "informatikk": [1, 2, 3, 4, ..., 57, 58, 59, 60],
    "gøy": [1, 59]
}
```

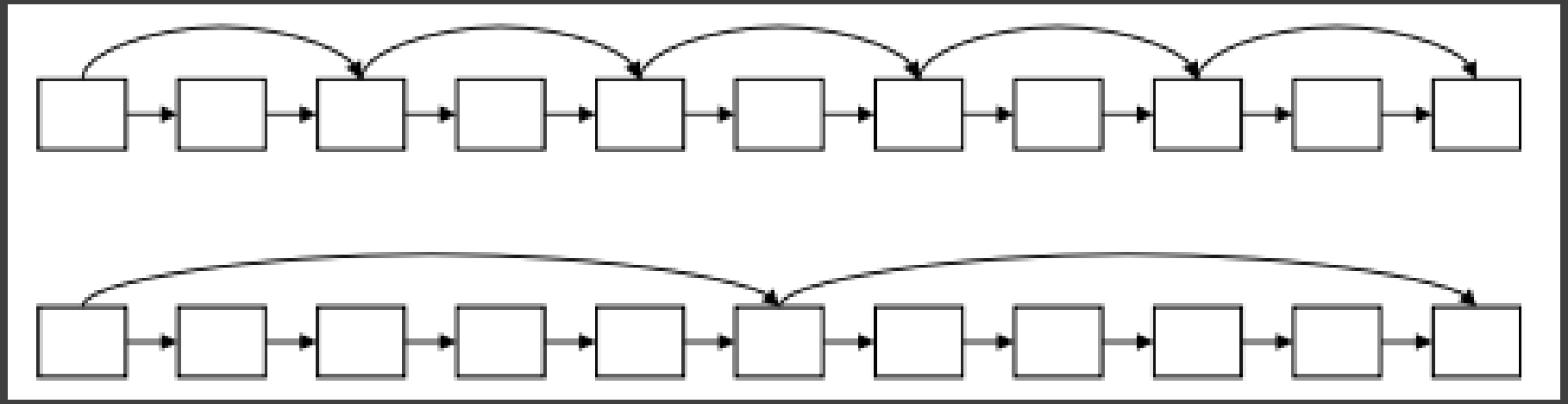
Skip pointers

- Eksempel: Hvis begge pekerne er på 8



Skip pointers

- Hvor plassere skips?
  - Flere skips: Mer sannsynlig å hoppe, men mer å sammenligne
  - Færre skips: Mindre sannsynlig å hoppe, men mindre å sammenligne
- Forslag: Med lengde L, fordel pointers på hver $\sqrt{L}$ posting



Agenda

Første time

- Science fair
- Repetisjon
- Mer om Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing

Assignment A

- Innleveringer
- Tips og triks til testene
- Iter-metoder
- Hva er fields i et dokument?

Levering

- Devilry
- Kan levere i par
- Gjerne lever .zip eller filene (jeg foretrekker bare filer)
- Lever bare filene som står i oppgaveteksten

Tester

- Hvordan kjøre testene (gjelder alle obliger)
- Printing fra filene
- Eksempel: Hvordan ser egentlig document-objekter ut?

assignments.py

- Alle obliger har en test suite
- Kan kommentere bort filer for å dele opp testing

```
def assignment_a_suite() -> unittest.TestSuite:
    # return build_test_suite(["TestInMemoryInvertedIndexWithoutCompression",
    #                          "TestPostingsMerger", "TestBooleanSearchEngine"])
    return build_test_suite(["TestInMemoryInvertedIndexWithoutCompression"])
```

Tester i invertedindex

```python
def test_access_postings(self):
    corpus = in3120.InMemoryCorpus()
    corpus.add_document(in3120.InMemoryDocument(0, {"body": "this is a Test"}))
    corpus.add_document(in3120.InMemoryDocument(1, {"body": "test TEST prøve"}))
    index = in3120.InMemoryInvertedIndex(corpus, ["body"], self._normalizer, self._tokenizer, self._compressed)
    assert 1
    # self.assertListEqual(list(index.get_terms("PRøvE wtf tesT")), ["prøve", "wtf", "test"])
    # self.assertListEqual([(p.document_id, p.term_frequency) for p in index["prøve"]], [(1, 1)])
    # self.assertListEqual([(p.document_id, p.term_frequency) for p in index.get_postings_iterator("wtf")], [])
    # self.assertListEqual([(p.document_id, p.term_frequency) for p in index["test"]], [(0, 1), (1, 2)])
    # self.assertEqual(index.get_document_frequency("wtf"), 0)
    # self.assertEqual(index.get_document_frequency("prøve"), 1)
    # self.assertEqual(index.get_document_frequency("test"), 2)
    # self.assertEqual(index.get_collection_frequency("wtf"), 0)
    # self.assertEqual(index.get_collection_frequency("prøve"), 1)
    # self.assertEqual(index.get_collection_frequency("test"), 3)
```

```python
def test_multiple_fields(self):
    document = in3120.InMemoryDocument(0, {
        'felt1': 'Dette er en test. Test, sa jeg. TEST!',
        'felt2': 'test er det',
        'felt3': 'test TEsT',
    })
    corpus = in3120.InMemoryCorpus()
    corpus.add_document(document)
    index = in3120.InMemoryInvertedIndex(corpus, ['felt1', 'felt3'], self._normalizer, self._tokenizer, self._compressed)
    assert 1
    # posting = next(index.get_postings_iterator('test'))
    # self.assertEqual(posting.document_id, 0)
    # self.assertEqual(posting.term_frequency, 5)
```

Iterasjon over Corpus

- Corpus-objektet inneholder en liste med document-objekter

build_index

```
for document in self._corpus:
            print(document)
```

```
test_access_postings (test_inmemoryinvertedindexwithoutcompression.TestInMemoryInvertedIndexWithoutCompression.test_access_postings) ...
{'document_id': 0, 'fields': {'body': 'this is a Test'}}
{'document_id': 1, 'fields': {'body': 'test TEST prØve'}}
ok
test_multiple_fields (test_inmemoryinvertedindexwithoutcompression.TestInMemoryInvertedIndexWithoutCompression.test_multiple_fields) ...
{'document_id': 0, 'fields': {'felt1': 'Dette er en test. Test, sa jeg. TEST!', 'felt2': 'test er det', 'felt3': 'test TEsT'}}
ok
```

Fields

- Et dokument består ofte av flere seksjoner
  - Header, body, footer, ...
- Blir ikke nødvendigvis behandlet likt
- Vi vil gå gjennom alle fields vi får som parametere

```
test_access_postings (test_inmemoryinvertedindexwithoutcompression.TestInMemoryInvertedIndexWithoutCompression.test_access_postings) ...
{'document_id': 0, 'fields': {'body': 'this is a Test'}}
{'document_id': 1, 'fields': {'body': 'test TEST prØve'}}
ok
test_multiple_fields (test_inmemoryinvertedindexwithoutcompression.TestInMemoryInvertedIndexWithoutCompression.test_multiple_fields) ...
{'document_id': 0, 'fields': {'felt1': 'Dette er en test. Test, sa jeg. TEST!', 'felt2': 'test er det', 'felt3': 'test TEsT'}}
ok
```

Agenda

Første time

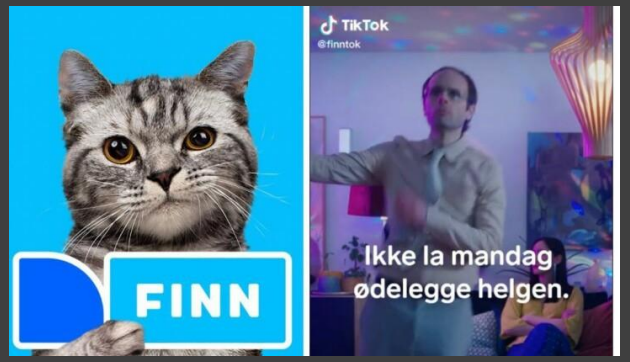
- Science fair
- Repetisjon
- Mer om Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing

Ukas Shoutout: Finn.no sine reklamer

- Reiseleder-Kjell
- Squeeze day
- Mandag
- …



Agenda

Første time

- Science fair
- Repetisjon
- Mer om Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing

Selvstendig jobbing!

Pause 15 min