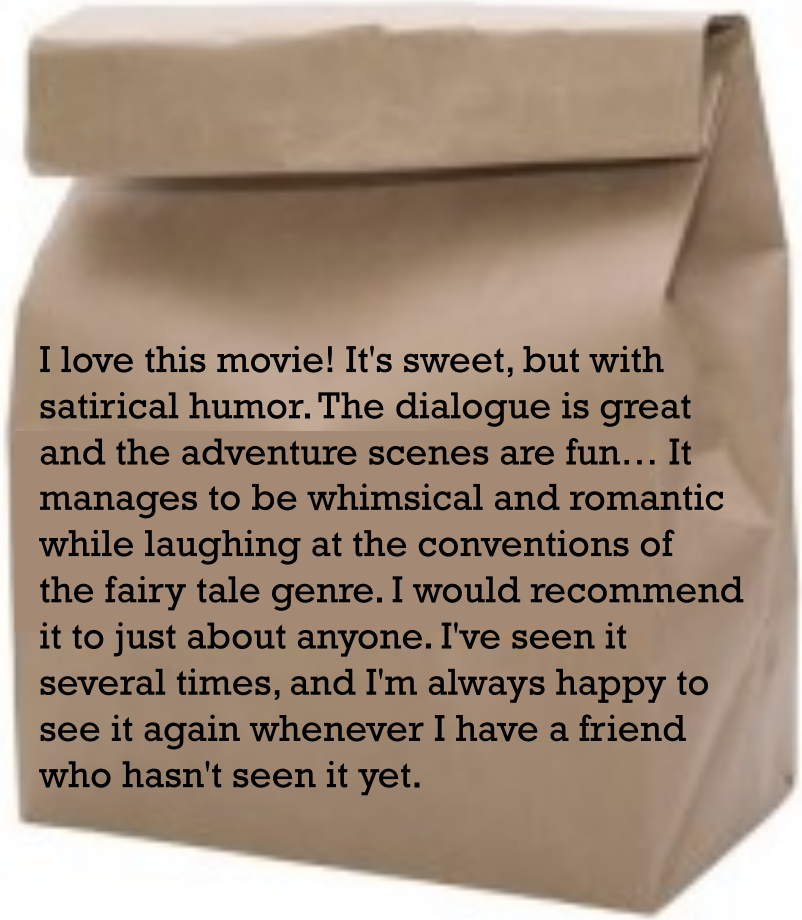
IMPROVING TEXT CLASSIFICATION

BAG OF WORDS

- **Bag of Words**: Assume position doesn't matter
- **Conditional Independence**: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c.

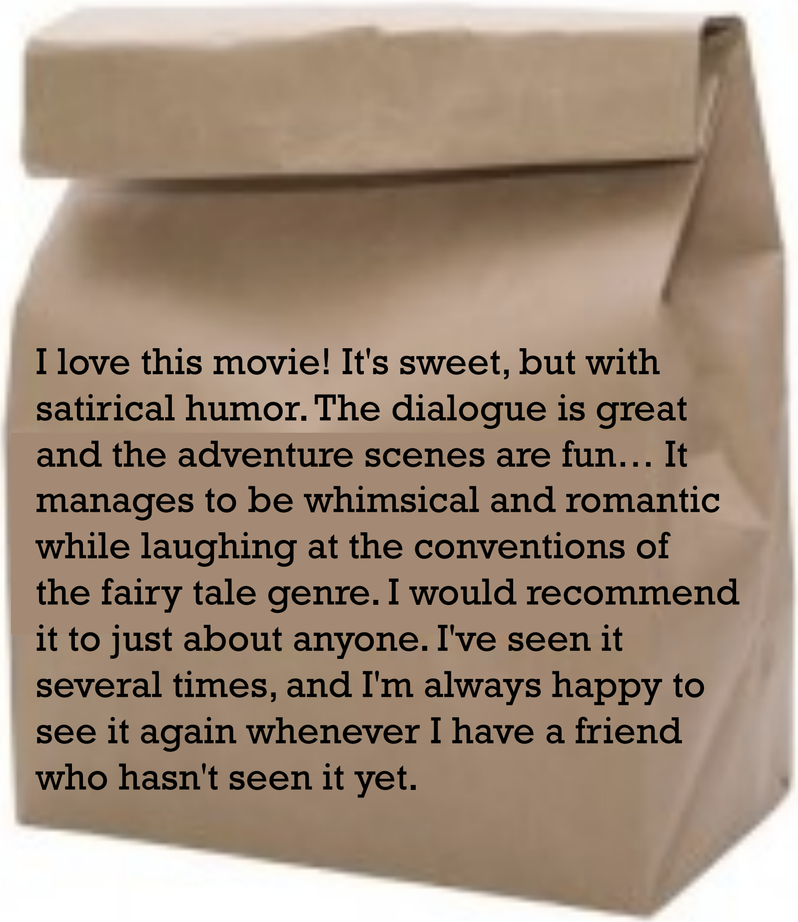


ADVANTAGES

- Very Fast, low storage requirements
- Robust to Irrelevant Features
  - Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features
  - Decision Trees suffer from *fragmentation* in such cases – especially if little data
- Optimal if the independence assumptions hold:
  - If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification

POSSIBLE IMPROVEMENTS

- Stop words
- Stemming
- TF-IDF
- Bigrams
- More



$$\gamma\left(\begin{array}{ll} \textbf{it} & 6 \\ \textbf{I} & 5 \\ \textbf{the} & 4 \\ \textbf{to} & 3 \\ \textbf{and} & 3 \\ \text{seen} & 2 \\ \text{yet} & 1 \\ \text{love} & 1 \\ \text{movie} & 1 \\ \text{sweet} & 1 \\ \ldots & \end{array}\right) = c$$

STOP WORDS

- Commonly used words
- May not contribute to how text is classified
- Possible wasted time and space
- Making stop word list
  - Collect most frequent terms
  - Keep those with relevant semantic content
- Pre-process all documents to remove stop words

a
about
above
after
again
against
all
am
an
and
any
are
aren't
as
at
be
because
been
before
being
below
between
both
but
by
can't
cannot
could
couldn't
did
didn't
do
does
doesn't
doing
don’t
down
during
each
few
for
from
further
had
hadn't
has

car
cars

movie
movies

laugh
laughs
laughing
laughed
laughter

replace
replaces
replaced
replacement

swim
swims
swam
swimming
swimmer

hope
hopes
hoped
hoping
hopeful
hopefully
hopefulness
hopeless
hopelessly
hopelessness

pony
ponies
ponied

quick
quicker
quickest
quickly

hero
heroes
heroic
heroically

STEMMING

- Many forms of a word may be used
- Stemming: Chop off the end of words
- Crude heuristic process
- Porter's algorithm
  - 5 step process
  - Changes ending based on some conditions

- Example rules:

sses → ss	possesses → possess
ies → i	movies → movi
ss → ss	possess → possess
s →	cars → car
(m>1) ement →	replacement → replac
	cement → cement

TERM FREQUENCY

- d_1: This is the best movie of the year.
- d_2: This movie features a great story with a great cast.
- "this"
  - Raw count:
    - f("this",d_1)=1
    - f("this",d_2)=1
  - Frequency:
    - tf("this",d_1)=1/8
    - tf("this",d_2)=1/10

Term	Term count
this	1
is	1
the	2
best	1
movie	1
of	1
year	1

Term	Term count
this	1
a	2
features	1
great	2
movie	1
with	1
story	1
cast	1

INVERSE DOCUMENT FREQUENCY

- A measure of how much information a word provides
  - If a word is in most (or all) documents, it doesn't help tell us how to classify the document
- Ratio of how many documents in corpus and how many documents with a given word

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

- "this"
  - idf("this", D) = log 2/2 = 0
- "best"
  - idf("best",D) = log 2/1 = 0.301

Term	Term count
this	1
is	1
the	2
best	1
movie	1
of	1
year	1

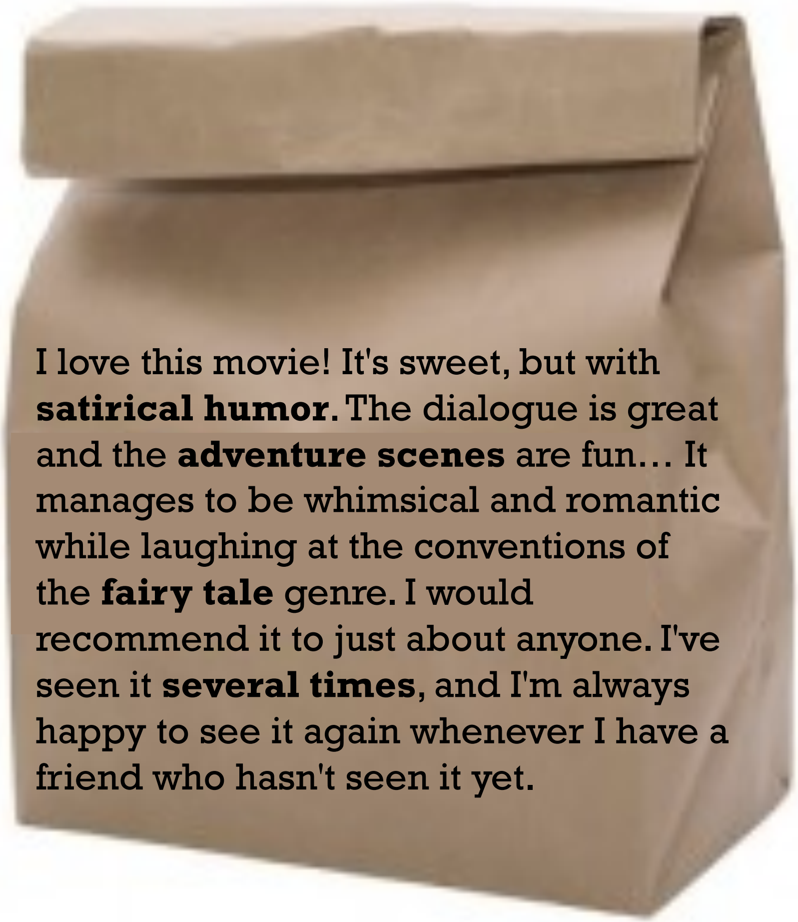
Term	Term count
this	1
a	2
features	1
great	2
movie	1
with	1
story	1
cast	1

TF-IDF

- Term frequency – inverse document frequency
- tfidf(t,d) = tf(t,d) * idf(t,D)
- "this"
  - tfidf("this",d_1) = tf("this",d_1) * idf("this",D)
    = 0.125 * 0 = 0
  - tfidf("this",d_2) = tf("this",d_2) * idf("this",D)
    = 0.1 * 0 = 0
- "best"
  - tfidf("best",d_1) = tf("best",d_1) * idf("best",D)
    = 0.125 * 0.301 = 0.038
  - tfidf("best",d_2) = tf("best",d_2) * idf("best",D)
    = 0 * 0.301 = 0

Term	Term count
this	1
is	1
the	2
best	1
movie	1
of	1
year	1

Term	Term count
this	1
a	2
features	1
great	2
movie	1
with	1
story	1
cast	1



BIGRAMS

$$\gamma\left(\begin{array}{ll} \text{this movie} & 6 \\ \text{I have} & 3 \\ \text{of the} & 7 \\ \textbf{fairy tale} & 2 \\ \text{and the} & 8 \\ \textbf{satirical humor} & 1 \\ \textbf{adventure scenes} & 1 \\ \textbf{several times} & 1 \\ \text{seen it} & 3 \\ \text{a friend} & 4 \\ \ldots & \end{array}\right) = c$$

MORE

- Capitalization
- Punctuation
- Contractions
- Misspellings
- Jargon