Tweets.csv File

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	tweet_id	airline	text										
2	1	Virgin Ameri	@VirginAmerica What @dhepburn said.										
3	2	Virgin Ameri	@VirginAmerica plus you've added commercials to the experience... tacky.										
4	3	Virgin Ameri	@VirginAmerica I didn't today... Must mean I need to take another trip!										
5	4	Virgin Ameri	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse										
6	5	Virgin Ameri	@VirginAmerica and it's a really big bad thing about it										
7	6	Virgin Ameri	@VirginAm										
8	7	Virgin Ameri	@VirginAmerica yes, nearly every time I fly VX this ‚Äúear worm‚Äù won‚Äôt go away :)										
9	8	Virgin Ameri	@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP										
10	9	Virgin Ameri	@virginamerica Well, I didn't‚Äîbut NOW I DO! :-D										
11	10	Virgin Ameri	@VirginAmerica it was amazing, and arrived an hour early. You're too good to me.										
12	11	Virgin Ameri	@VirginAmerica did you know that suicide is the second leading cause of death among teen										
13	12	Virgin Ameri	@VirginAmerica I <3 pretty graphics. so much better than minimal iconography. :D										
14	13	Virgin Ameri	@VirginAmerica This is such a great deal! Already thinking about my 2nd trip to @Australia										

Tweets.csv

```
tweet_id,airline,text
1,Virgin America,@VirginAmerica What @dhepburn said.
2,Virgin America,@VirginAmerica plus you've added commercials to the experience... tacky.
3,Virgin America,@VirginAmerica I didn't today... Must mean I need to take another trip!
4,Virgin America,"@VirginAmerica it's really aggressive to blast obnoxious
""entertainment"" in your guests' faces &amp; they have little recourse"
5,Virgin America,@VirginAmerica and it's a really big bad thing about it
6,Virgin America,"@VirginAmerica seriously would pay $30 a flight for seats that didn't
have this playing.
it's really the only bad thing about flying VA"
7,Virgin America,"@VirginAmerica yes, nearly every time I fly VX this “ear worm” won’t go
away :)"
8,Virgin America,"@VirginAmerica Really missed a prime opportunity for Men Without Hats
parody, there. https://t.co/mWpG7grEZP"
9,Virgin America,"@virginamerica Well, I didn't…but NOW I DO! :-D"
10,Virgin America,"@VirginAmerica it was amazing, and arrived an hour early. You're too
good to me."
11,Virgin America,@VirginAmerica did you know that suicide is the second leading cause of
death among teens 10-24
12,Virgin America,@VirginAmerica I &lt;3 pretty graphics. so much better than minimal
iconography. :D
13,Virgin America,@VirginAmerica This is such a great deal! Already thinking about my 2nd
trip to @Australia &amp; I haven't even gone on my 1st trip yet! ;p
14,Virgin America,@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies
again! U take all the #stress away from travel http://t.co/ahlXHhKiyn
15,Virgin America,@VirginAmerica Thanks!
```

Load Tweets Data

```
DROP TABLE IF EXISTS tweets_raw;
CREATE EXTERNAL TABLE tweets_raw (
  tweet_id int,
  airline string,
  text string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS
INPUTFORMAT
'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION '${INPUT}/tweets/'
TBLPROPERTIES (
  "s3select.format" = "csv",
  "s3select.headerInfo" = "ignore"
);
```

1. If table already exists, delete it
2. Create a table named “tweets_raw” with
   an integer column named “tweet_id”,
   a string column named “airline”, and
   a string column named “text”

Since we load from csv file, it’s comma separated

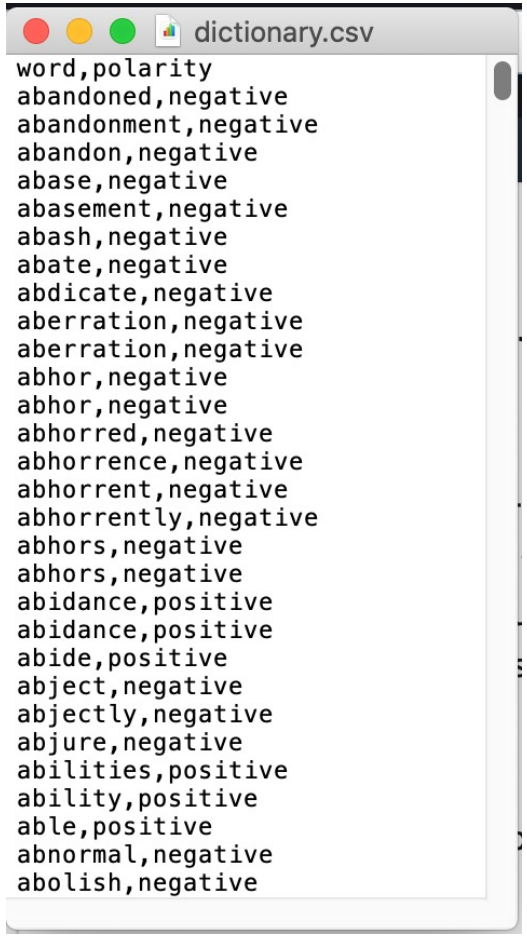
We use pre-defined input and output formats, defined by AWS and Apache

This is where we read original data. ${INPUT} is the path you point to at adding step, usually *s3://yourBuckect/input/*

We specify table properties: how to select data from S3 objects, and ignore the first line of csv which is the header

Dictionary.csv File

	A	B
1	word	polarity
2	abandoned	negative
3	abandonmer	negative
4	abandon	negative
5	abase	negative
6	abasement	negative
7	abash	negative
8	abate	negative
9	abdicate	negative
10	aberration	negative
11	aberration	negative
12	abhor	negative
13	abhor	negative
14	abhorred	negative
15	abhorrence	negative



Load Dictionary Data

```
DROP TABLE IF EXISTS dictionary;
CREATE EXTERNAL TABLE dictionary (
  word string,
  polarity string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS
INPUTFORMAT
'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION '${INPUT}/dictionary/'
TBLPROPERTIES (
  "s3select.format" = "csv",
  "s3select.headerInfo" = "ignore"
);
```

1. If table already exists, delete it
2. Create a table named “dictionary” with
   a string column named “word”, and
   a string column named “polarity”

Since we load from csv file, it’s comma separated

We use pre-defined input and output formats, defined by AWS and Apache

This is where we read original data. ${INPUT} is the path you point to at adding step, usually *s3://yourBuckect/input/*

We specify table properties: how to select data from S3 objects, and ignore the first line of csv which is the header

Tokenization

```
drop view if exists l1;

create view l1 as select tweet_id, words from tweets_raw lateral view
explode(sentences(lower(text))) dummy as words;

drop view if exists l2;

create view l2 as select tweet_id, word from l1 lateral view explode( words ) dummy as word;
```

1. A view is a virtual table. Views are definitions built on top of other tables (or views), and do not hold data themselves, like taking a snapshot
2. Explode each tweet text in array of words.
   explode() takes in an array as an input and outputs the elements of the array as separate rows
   Lateral view creates a virtual table for exploded columns and make join with the original table
   A dummy table is a virtual table. In Oracle, the name is DUAL

4. Explode each tweet -> word on multiple rows

Feature Extraction

```
drop view if exists l3;
create view l3 as select
  tweet_id,
  l2.word,
  case d.polarity
    when  'negative' then -1
    when 'positive' then 1
    else 0 end as polarity
from l2 left outer join dictionary d on
l2.word = d.word;
```

Left outer join l2 with dictionary on matching word

case is conditional statement in the format:

CASE

WHEN condition_1 THEN result_1

WHEN condition_2 THEN result_2

ELSE result_n

END

If the word has polarity “negative” in dictionary table, use value -1

If the word has polarity “positive” in dictionary table, use value 1

If the word has polarity “neutral” in dictionary table, use value 0

Note that NULL entry will also be replaced by 0

the value -1, 1 or 0 will be stored in a column called “polarity”

Classification

```
DROP TABLE IF EXISTS tweets_sentiment;
create table tweets_sentiment stored as orc as select
  tweet_id,
  case
  when sum( polarity ) > 0 then 'positive'
  when sum( polarity ) < 0 then 'negative'
  else 'neutral' end as sentiment
from l3 group by tweet_id;
```

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}' SELECT * FROM
tweets_sentiment;
```

1. Create table named “tweets_sentiment” from l3, and group by tweet_id, i.e. merge all rows with same tweet_id together

The Optimized Row Columnar (ORC) file format, a highly efficient way to store Hive data, a single file as the output of each task

Sum up all polarity (-1/1/0) already grouped by tweet_id
If sum is larger than 0, uses “positive”
If sum is less than 0, uses “negative”
If sum is 0, uses “neutral”
this value is stored in a column called “sentiment”

2. Export the data to output directory
${OUTPUT} is the path you point to at adding step, usually *s3://yourBuckect/output/* which does not exist yet