Project Big Data
Sentiments Analysis

Objective:
The objective of this project is to perform **Sentiment Analysis** using the MapReduce programming model. You will analyze a dataset of IMDB reviews and classify the sentiment (positive, negative, neutral) based on the presence of predefined keywords.

Prerequisites:

- Basic understanding of the MapReduce framework.
- Python installed (if using Hadoop streaming) or access to a Hadoop cluster.

Dataset Example

We will use the **IMDB_Review_Dataset.csv**, which contains customer reviews of various movies.

- Each line in the dataset represents a single review.
- Example

"One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me.

The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word.

It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away.

I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side."
"A wonderful little production.

The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece.

The actors are extremely well chosen- Michael Sheen not only ""has got all the polari"" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrificly written and performed piece. A masterful production about one of the great master's of comedy and his life.

The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done.",positive
"I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue is witty and the characters are likable (even the well bread suspected serial killer). While some may be disappointed when they realize this is not Match Point 2: Risk Addiction, I thought it was proof that Woody Allen is still fully in control of the style many of us have grown to love.

This was the most I'd laughed at one of Woody's comedies in years (dare I say a decade?). While I've never been impressed with Scarlet Johanson, in this she managed to tone down her ""sexy"" image and jumped right into a average, but spirited young woman.

This may not be the crown jewel of his career, but it was wittier than ""Devil Wears Prada"" and more interesting than ""Superman"" a great comedy to go see with friends.",positive
"Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.

This movie is slower than a soap opera... and suddenly, Jake decides to become Rambo and kill the zombie.

OK, first of all when you're going to make a film you must Decide if its a thriller or a drama! As a drama the movie is watchable. Parents are divorcing & arguing like in real life. And then we have Jake with his closet which totally ruins all the film! I expected to see a BOOGEYMAN similar movie, and instead i watched a drama with some meaningless thriller spots.

3 out of 10 just for the well playing parents & descent dialogs. As for the shots with Jake: just ignore them."
"Petter Mattei's ""Love in the Time of Money"" is a visually stunning film to watch. Mr. Mattei offers us a vivid portrait about human relations. This is a movie that seems to be telling us what money, power and success do to people in the different situations we encounter.

This being a variation on the Arthur Schnitzler's play about the same theme, the director transfers the action to the present time New York where all these different characters meet and connect. Each one is connected in one way, or another to the next person, but no one seems to know the previous point of contact. Stylishly, the film has a sophisticated luxurious look. We are taken to see how these people live and the world they live in their own habitat.

The only thing one gets out of all these souls in the picture is the different stages of loneliness each one inhabits. A big city is not exactly the best place in which human relations find sincere fulfillment, as one discerns is the case with most of the people we encounter.

The acting is good under Mr. Mattei's direction. Steve Buscemi, Rosario Dawson, Carol Kane, Michael Imperioli, Adrian Grenier, and the rest of the talented cast, make these characters come alive.

We wish Mr. Mattei good luck and await anxiously for his next work."
"Probably my all-time favorite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy or boring. It just never gets old, despite my having seen it some 15 or more times in the last 25 years. Paul Lukas' performance brings tears to my eyes, and Bette Davis, in one of her very few truly sympathetic roles, is a delight. The kids are, as grandma says, more like ""dressed-up midgets"" than children, but that only makes them more fun to watch. And the mother's slow awakening to what's happening in the world and under her own roof is believable and startling. If I had a dozen thumbs, they'd all be ""up"" for this movie."

Tasks

1. Collect the Sentiment Keywords Dictionary

- **Positive Keywords**:
  Examples include *love, amazing, happy, good, excellent, wonderful, fantastic, brilliant, satisfied, success*.
- **Negative Keywords**:
  Examples include *horrible, frustrating, terrible, bad, worst, disappointing, awful, regret*.

2. Preprocess the Dataset

- Explore preprocessing techniques to clean the data:
  - Remove irrelevant information, such as **stop words** (e.g., *the, is, and, or*).
  - Strip special characters and punctuation to standardize the text.

3. Create the MapReduce Program

- Implement a program to process the dataset and classify reviews based on the sentiment dictionary.

4. Upload Dataset to HDFS

- Load the IMDB_Review_Dataset.csv file into the Hadoop Distributed File System (HDFS).

5. Run MapReduce Job

- Execute the MapReduce program to classify reviews as **Positive**, **Negative**, or **Neutral**.

6. Fetch Results

- Collect the sentiment analysis results from the HDFS output.

7. Display Results in a Graph

- Use visualization tools (e.g., Matplotlib, Excel, or Tableau) to create a bar chart or pie chart showing the count of positive, negative, and neutral reviews.

Analysis Questions

1. **Which sentiment dominates the dataset?**
   Analyze the visualized results to determine whether the majority of reviews are positive, negative, or neutral.
2. **What does this imply about the overall sentiment in the IMDB reviews?**
   Reflect on the broader implications of your findings (Consumer Perception, Platform Trustworthiness, Economic Insights).