

Natural Language Processing - Assignment 1:

Classical Baselines + Evaluation Pack

Zofia Drózdź (S5853214), Lucía Adán Castejón (S5908671), Aina Seguer Estape
(S5934524)

1 Dataset & Setup

The dataset is already split into two categories: test, which has 7.5k rows, and train, which has 120k rows. For our model, however, we want to use 3 categories: train, test, and dev. The dev and the final train set are based on the initial train set; therefore, we used the function `train_test_split(temp_train_set, test_size=0.1, random_state=42, stratify=temp_train_set["label"])` from the library `sklearn`. We used the random seed 42, in order to allow the results to be reproducible. This function split the official train set from the data into a training set (90%) and a dev set, also referred to as a validation set (10%). The random seed of 42 was used throughout the assignment for reproducibility.

The next step of the preprocessing is to combine the title and description for all the samples, as it is not important to differentiate the words being in the description or in the title. We store the labels as `y` and the title and description as `ag`. The last step of preprocessing is to implement the word-level TF-IDF. This counts the number of times each word appears in the dataset and also considers the importance of the words in the sentences; the `TfidfVectorizer` from `sklearn` outputs a sparse vector containing this information. The vectors are stored as `X`, and this is done for the 3 sets: training, validation, and test. Moreover, this function makes all the letters lowercase and removes the stop-words, which are, for example, "the", "a", etc. These are words that don't add a lot of meaning, and are frequent in English. TF-IDF also determines the maximum number of features, which we set up as 5000 as was done in the tutorial notebooks. With this, the vectorizer will build a vocabulary consisting of the top 5000 words that appear most frequently in the data. The `TfidfVectorizer` also performs tokenisation, which is breaking down text into smaller tokens made up of words, and turns all words into lowercase (3). TF-IDF is performed on the title and the description.

2 Models

The logistic regression model we trained was the `LogisticRegression` model from the `sklearn.linear_model` library. The model was trained using the `saga` solver, which stands for "stochastic average gradient descent". We chose this solver because it has been found to work well for large datasets and features with large dimensionality (1), which is the case for our model since the training set has 108 000 samples and with TF-IDF the dimensionality amounts to the number of words in the vocabulary or the `max_features` parameter which we set to 5000 as was done in the tutorial notebook. For this solver, in a multi-class setting, the solver will try to optimize the multinomial loss.

The linear SVM model we trained the `LinearSVC` model from the `sklearn.svm` library. In the multiclass setting, this model works in a one-vs-rest scheme, training 4 one-vs-rest classifiers. This model is based on the `liblinear` solver, and has more flexibility in loss and penalty choices, as well as scaling better to larger numbers of samples (2).

3 Experimental Protocol

Our experimental protocol was made up of a grid search to tune hyper-parameters, resulting in a selection of the best for each of the two baseline models according to the Macro-F1 metric on the dev set. We made the grid search manually as opposed to using `GridSearchCV` from `sklearn` as we wanted to use our dev set for tuning rather than doing k-fold cross-validation due to the high number of training samples and computational time required.

For the logistic regression, the hyper-parameters we chose to tune were `C` and `l1_ratio`, which control regularization. For `C` we tested values 0.1 and 1 as these are typical good choices to control regularization strength. For `l1_ratio` we tested 0 and 1 to apply only l2 or only l1 respectively. For the linear SVM, the hyper-parameters we chose to tune were the `loss` being optimized and the regularization parameter `C`. We tested both the `hinge` and `squared_hinge` losses as these are the available ones, and the same values for `C`. We originally wished to try out more values and tune further hyper-parameters, but this resulted in run times exceeding 90 minutes which is why we had to cut down.

To select the best model from the grid search we used the Macro-F1 metric, as it is our secondary metric but takes into account the types of errors being made more than our primary metric of accuracy. For the logistic regression, the best hyper-parameters we found were `C=1`, `l1_ratio=0`, and they resulted in a Macro-F1 of 0.911. The best hyper-parameters for the linear SVM were `loss="hinge"`, `C=1` and they resulted in a Macro-F1 of 0.914.

4 Results

With our best models selected, we evaluate them on the validation set first and then on the test set. We use the classification report and confusion matrix from the `sklearn.metrics` library.

Logistic Regression

Category	P	R	F1	Sup
World	0.92	0.90	0.91	3000
Sports	0.95	0.98	0.96	3000
Business	0.88	0.88	0.88	3000
Sci/Tech	0.89	0.90	0.89	3000
Accuracy			0.91	12k
Macro F1	0.91	0.91	0.91	12k

Category	P	R	F1	Sup
World	0.92	0.90	0.91	1900
Sports	0.95	0.98	0.96	1900
Business	0.87	0.87	0.87	1900
Sci/Tech	0.88	0.88	0.88	1900
Accuracy			0.90	7600
Macro F1	0.90	0.90	0.90	7600

Table 1: Classification Report of Logistic Regression on Validation (left) and Test set (right)

Validation Set: Reading from the Table 1, we get that the model gets very high Accuracy, Precision, Recall and f1 score. Moreover, the macro f1 score is also high - 0.91, which is the same as the accuracy, thus our primary and secondary metrics both indicate this is a good model. The confusion matrix shows that the class which is classified correctly the most and the least are *Sports* and *Business*, respectively. The classes which are misclassified the most as the other are *Sci/Tech* and *Business*, and the class which is generally misclassified the least is *Sports*. These observations suggest that classes such as *Business* and *Sci/Tech* have higher overlap or more similarities compared to others.

Test Set: From the Table 1 we get that, similarly to the validation set, the model performs well. However, the accuracy and macro f1 are minimally lower - both 0.90. The difference is not significant enough to be considered distressing. It is possibly caused by the change in the size of the set. The confusion matrix is very similar to the one of the validation set - the conclusions do not change. Overall, the results present similar high performance on the test data.



(a) Confusion Matrix of Logistic Regression on the validation set

(b) Confusion Matrix of Logistic Regression on the test set

Figure 1: Confusion Matrices for Logistic Regression

Linear SVM

Validation Set: From the Table 2 we can read that the SVM model gets quite high results for Accuracy, Precision, Recall, and f1 score across the four classes. The accuracy and macro f1 specifically are also

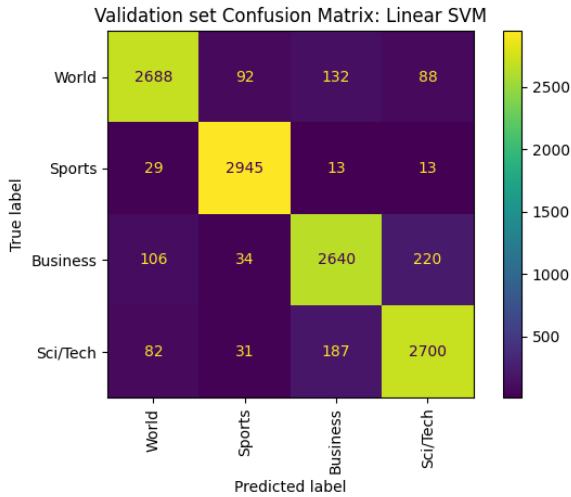
Category	P	R	F1	Sup
World	0.93	0.90	0.91	3000
Sports	0.95	0.98	0.97	3000
Business	0.89	0.88	0.88	3000
Sci/Tech	0.89	0.90	0.89	3000
Accuracy			0.91	12k
Macro F1	0.91	0.91	0.91	12k

Category	P	R	F1	Sup
World	0.93	0.90	0.91	1900
Sports	0.95	0.98	0.97	1900
Business	0.87	0.87	0.87	1900
Sci/Tech	0.88	0.88	0.88	1900
Accuracy			0.91	7600
Macro F1	0.91	0.91	0.91	7600

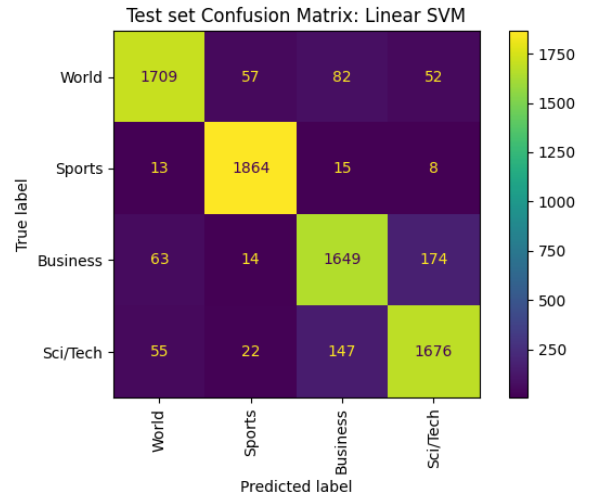
Table 2: Classification Report of Linear SVM on Validation (left) and Test set (right)

high - both 0.91. Looking at the confusion matrix, we can see that the same behavior as in the logistic regression is present. The confusion matrices are almost identical in terms of the most/least true positives and false positives/negatives to the Logistic Regression.

Test Set: From the classification report, Table 2, we get that the the model performs almost identically as on the validation set. There are only slight differences (lower values) in the results from the *Business* and *Sci/Tech* classes. Nevertheless, the metrics still suggests very good performance of the model overall - it is adept at predicting the labels of the articles. From the confusion matrix we can derive similar conclusions as from the validation set evaluation. Overall, the test set bears similar results to the validation set.



(a) Confusion Matrix of SVM on the validation set



(b) Confusion Matrix of SVM on the test set

Figure 2: Confusion Matrices for SVM

Comparison of the models

From the results in the classification reports, we can conclude that the SVM model performs marginally better than the Logistic Regression model, although there is a small difference. This is mainly deduced from the classification reports (Accuracy, Precision, Recall, macro f1 score). Considering the confusion matrices, we also see the same tendencies of how labels are assigned. However, we also observe that the SVM on the validation set shows less true positives in the *World* class, and consequently more incorrect classifications for that category. Other classes present better values than the Logistic Regression. Regarding the SVM on the test set, only the incorrect classification of *Business* as the *World* class is higher than in the Logistic Regression.

5 Error Analysis

For the error analysis, we collected 20 out of the total misclassified examples from the test set for each of the models.

20 misclassified examples from the test set for the logistic regression model can be seen in Table 3 (in the Appendix section), from the total 731 errors out of 7600 predictions that the logistic regression made.

20 misclassified examples from the test set for the linear SVM can be seen in Table 4 (in the Appendix section), from the total 702 errors out of 7600 predictions that the SVM made.

It is interesting to point out that out of these 20 misclassified examples for each of the models, 18 of the first errors overlap between the two models. This suggests that these examples may be close to the decision boundary and that the two models may have similar decision boundaries.

As per our Results discussion, the linear SVM model performed better as it has a higher accuracy and Macro-F1 score on both the validation and test sets, which are our primary and secondary evaluation metrics respectively. Thus, for this model we will categorize 10 of the example misclassifications and provide some insights into why these examples were misclassified.

Overlap of the Labels

- **IBM to hire even more new workers** (Row 3)
 - The title and the description explain how IBM is hiring new workers. In this case, the labels overlap, as it is true that IBM is a technology company and therefore could fall under the Sci/Tech label. However, the fact that the article focuses on hiring and workforce expansion can also be understood as a business-related topic. Therefore, the labels overlap, and the article could reasonably be assigned to both categories. The human label of Sci/Tech is not incorrect, but the model is also not incorrect in predicting Business.
- **Some People Not Eligible to Get in on Google IPO** (Row 4)
 - The article discusses Google’s initial public offering and explains how, despite claims of openness, certain groups are excluded from participating in the IPO process. Although Google is a technology company, the main focus of the article is on financial mechanisms, investment access, and market participation. For this reason, the article can be seen as belonging to the Business category. At the same time, its association with a major technology firm makes a Sci/Tech label understandable. As a result, this represents another case of overlapping labels rather than a clear misclassification.
- **Intel to delay product aimed for high-definition TVs** (Row 8)
 - The article discusses Intel’s latest delays in releasing products, specifically of a video display chip it had planned to introduce by the year end. This article talks about chips and other technological products, which is why the model labeled it as Sci/Tech. However, it is also true that Intel is a company and the article mentions the impacts of the delays on a fast-growing market, which is why the true label was Business. This is another case where the article really falls into two categories at once, thus appearing to be misclassified.
- **Yahoo! Ups Ante for Small Businesses** (Row 11)
 - This article is about how the tech company "Yahoo!" is gambling that price cuts on its domain registration will make it more competitive. The real label on this was Business due to its focus on "Yahoo!"’s competitiveness in the market. However, the model classified it as Tech/Sci, which is not wrong as "Yahoo!" is indeed a tech company and the article discusses technological aspects relating to the the price cuts.

In these cases, the articles could be attributed to more than one label as they refer to topics related to more than one label. The label given by the human refers to one of the possible labels while the model decides to give it a different label. Both labels are correct, as the article refers to both, as there is an overlap of the labels. However, it is classified as an error as they are not the same, even if both labels are correct. This type of error points out that when classifying text, the decision is not always binary as text can refer to more than one thing at once

Human Labeling Error

- **Live: Olympics day four** (Row 7)
 - The label given for this article is World; however, based on both the title and the description, the article focuses on Olympic competitions and medal opportunities in Athens. We imagine humans may have mislabeled this as World because of the focus on Athens, but the truth is that the article is about the Olympics which are a sports event. Therefore, the correct label should be Sports, which is what the model predicted.
- **Security scare as intruder dives in** (Row 14)
 - The ground truth label for this article was Sports, likely because the event being described occurred at the Athens Olympics. However, this is an incorrect labeling since based on the description, the article describes the event of an intruder and security at the Olympics. Such an article more accurately corresponds to the World category as predicted by the model, as it discusses matters of safety at an international event.
- **U.S. Misses Cut in Olympic 100 Free** (Row 9)
 - Similarly to the case above, the article is labeled as World; however, it discusses the performance of American swimming sprinters in an Olympic event. As a result, it should be categorized as Sports, which again matches the model's prediction.

In these cases, even though the titles and descriptions contain country names, the primary focus of the articles is on sporting events rather than international affairs. The model correctly identifies the sports-related nature of the content, but the ground-truth labels are incorrect. These errors can therefore be attributed to human labeling mistakes rather than model misclassification. This type of error highlights some difficulties of supervised ML, as they require us to label data, but we can make mistakes. These mistakes can be costly, as if the training data is erroneous, then the model's output on the test data will most likely make those same mistakes.

Lexical Ambiguity

- **Venezuela Prepares for Chavez Recall Vote** (Row 6)
 - This article is highly political, as it discusses a recall vote in Venezuela. However, its description includes words such as "fraud", "oil", and "market" which have strong semantic relations to Business, and often appear in articles in about Business. Because TF-IDF embeddings can only capture one global meaning, the model cannot realize that this article is political but has implications in the oil market. Thus, this example shows how words can have distinct meanings based on context which the model may not be able to capture.
- **Hurricane center's projection on Charley not far off, data show** (Row 20)
 - This article talks about the scientific prediction of Hurricane Charley by the National Hurricane Center. The true label is therefore Sci/Tech, however the model predicted Business. This is likely because the description includes words such as "forecast", "projection", and "data", which tend to be linked to business. However, in this article they refer to weather forecasting predictions, thus there is an extent of lexical ambiguity in these words, which the model did not capture.
- **Teenage T. rex's monster growth** (Row 2)
 - This article is about the Tyrannosaurus rex growth reaching a massive size after a growth spurt. The article is clearly Sci/Tech as it is a paleontology discovery about a dinosaur. However, the model labeled it as Business since the description includes words like "growth" and "size". These words are often used to describe stocks, and were likely encountered in such a scenario many times by the model, so the model learned their meaning in that context. However, the model fails to capture the ambiguity in the words, as they can also be used to describe changes of a dinosaur over years.

In these cases, the descriptions of the articles use words that can be used in various contexts with various meanings. Thus, these words can be ambiguous for the model. Because the word embeddings used can only capture one global meaning, the model is associating them to their most frequently occurring context. So, in these examples where ambiguous words are used in a less frequent context, the model struggles to understand this difference.

6 Appendix

AI tools were used solely for formatting and for spelling and grammar checking. All content, analysis, and conclusions were conceived and written by the authors. The code was also written by the authors, with support from the Jupyter notebooks provided during course tutorial sessions.

Instructions to run the code from the GitHub repository (<https://github.com/luciaadan/nlp.git>) for reproducibility:

- To run the models first clone the repository to a device using the following command in the terminal

```
git clone https://github.com/luciaadan/nlp.git
cd nlp
```

- The project was done in WSL/MacOS and it requires a virtual environment to be run. The necessary files are in the repository but it also needs to be activated on the device so the following command must be executed in the terminal:

```
uv sync
source .venv/bin/activate
```

- Since the most important file is a python file, the code can be run in the terminal by the following commands:

```
cd notebooks
python3 assignment_1.py
```

Some other notes for reproducibility:

- The exact AG News dataset used is version 3, which was last updated 09/09/2015.
- Used official train/test split from HuggingFace. Official train set was split into 90/10 for train and dev sets, respectively.
- Random seed used: 42.
- Hyper-parameters for final logistic regression model:

```
- solver="saga"
- l1_ratio=0
- C=1
- max_iter=100
- random_state=42
- Note: All hyper-parameters not mentioned are used at their default value
```

- Hyper-parameters for final linear SVM model:

```
- C=1
- loss="hinge"
- max_iter=100
- random_state=42
- Note: All hyper-parameters not mentioned are used at their default value
```

- Approximate run time: 5 minutes

References

- [1] GeeksforGeeks. Comparing various online solvers in scikit learn. <https://www.geeksforgeeks.org/python/comparing-various-online-solvers-in-scikit-learn/>, 2025. Accessed: 19-Feb-2026.
- [2] scikit-learn developers. Linearsvc. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>, 2025. Accessed: 19-Feb-2026.
- [3] scikit-learn developers. Tfidfvectorizer. [://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.T](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

Table 3: Logistic Regression - 20 Misclassification Examples (Total Errors: 731)

Title	Description	True	Prediction
Prediction Unit Helps Forecast Wildfires (AP)	AP - It's barely dawn when Mike Fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts, but already he knows what the day will bring. Lightning will strike in places he expects. Winds will pick up, moist places will dry and flames will roar.	Sci/Tech	Sports
IBM to hire even more new workers	By the end of the year, the computing giant plans to have its biggest headcount since 1991.	Sci/Tech	Business
Some People Not Eligible to Get in on Google IPO	Google has billed its IPO as a way for everyday people to get in on the process, denying Wall Street the usual stranglehold it's had on IPOs. Public bidding, a minimum of just five shares, an open process with 28 underwriters - all this pointed to a new level of public participation. But this isn't the case.	Sci/Tech	Business
Rivals Try to Turn Tables on Charles Schwab	By MICHAEL LIEDTKE SAN FRANCISCO (AP) - With its low prices and iconoclastic attitude, discount stock broker Charles Schwab Corp. (SCH) represented an annoying stone in Wall Street's wing-tipped shoes for decades...	Sci/Tech	Business
Venezuela Prepares for Chavez Recall Vote	Supporters and rivals warn of possible fraud; government says Chavez's defeat could produce turmoil in world oil market.	World	Business
Live: Olympics day four	Richard Faulds and Stephen Parry are going for gold for Great Britain on day four in Athens.	World	Sports
Intel to delay product aimed for high-definition TVs	SAN FRANCISCO - In the latest of a series of product delays, Intel Corp. has postponed the launch of a video display chip it had previously planned to introduce by year end, putting off a showdown with Texas Instruments Inc. in the fast-growing market for high-definition television displays.	Business	Sci/Tech
U.S. Misses Cut in Olympic 100 Free	ATHENS, Greece - Top American sprinters Jason Lezak and Ian Crocker missed the cut in the Olympic 100-meter freestyle preliminaries Tuesday, a stunning blow for a country that had always done well in the event. Pieter van den Hoogenband of the Netherlands and Australian Ian Thorpe advanced to the evening semifinal a day after dueling teenager Michael Phelps in the 200 freestyle, won by Thorpe...	World	Sports
Stocks Climb on Drop in Consumer Prices	NEW YORK - Stocks rose for a second straight session Tuesday as a drop in consumer prices allowed investors to put aside worries about inflation, at least for the short term. With gasoline prices falling to eight-month lows, the Consumer Price Index registered a small drop in July, giving consumers a respite from soaring energy prices...	World	Business
Yahoo! Ups Ante for Small Businesses	Web giant Yahoo! is gambling that price cuts on its domain name registration and Web hosting products will make it more competitive with discounters in the space - which means that small businesses looking to move online get a sweeter deal through ...	Business	Sci/Tech
IBM Buys Two Danish Services Firms	IBM said Tuesday it has acquired a pair of Danish IT services firms as part of its effort to broaden its presence in Scandinavia. As a result of the moves, IBM will add about 3,700 IT staffers to its global head count. Financial terms of ...	Sci/Tech	Business
Oil prices bubble to record high	The price of oil has continued its sharp rise overnight, closing at a record high. The main contract in New York, light sweet crude for delivery next month, has closed at a record \$US46.75 a barrel - up 70 cents on yesterday's close.	World	Business
SUVs: Live And Let Die	NEW YORK - The newly released traffic crash fatality data have something for everyone in the debate about the safety of sport utility vehicles.	Business	Sci/Tech
Security scare as intruder dives in	A CANADIAN husband's love for his wife has led to a tightening of security at all Olympic venues in Athens.	Sports	World
Sharon orders 1,000 homes in West Bank	Israel announced plans for 1,000 houses in the West Bank yesterday, accelerating the expansion of the settlements.	World	Business
Google Lowers Its IPO Price Range	SAN JOSE, Calif. - In a sign that Google Inc.'s initial public offering isn't as popular as expected, the company lowered its estimated price range to between \$85 and \$95 per share, down from the earlier prediction of \$108 and \$135 per share...	World	Business
Future Doctors, Crossing Borders	Students at the Mount Sinai School of Medicine learn that diet and culture shape health in East Harlem.	World	Sci/Tech
Microsoft finalises three-year government deal	Hot on the heels of its 10-year strategic partnership with the London Borough of Newham, Microsoft is close to signing a new broad three-year public sector agreement with the government.	Sci/Tech	Business
Stock Prices Climb Ahead of Google IPO	NEW YORK - Investors shrugged off rising crude futures Wednesday to capture well-priced shares, sending the Nasdaq composite index up 1.6 percent ahead of Google Inc.'s much-anticipated initial public offering of stock. In afternoon trading, the Dow Jones industrial average gained 67.10, or 0.7 percent, to 10,039.93...	World	Business
Intuit Posts Wider Loss After Charge (Reuters)	Reuters - Intuit Inc. (INTU.O), maker of the No. 1 U.S. tax presentation software TurboTax, on Wednesday posted a wider quarterly loss after taking a goodwill impairment charge during its seasonally weaker fourth quarter.	Sci/Tech	Business

Table 4: Linear SVM - 20 Misclassification Examples (Total Errors: 702)

Title	Description	True	Prediction
Prediction Unit Helps Forecast Wildfires (AP)	AP - It's barely dawn when Mike Fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts, but already he knows what the day will bring. Lightning will strike in places he expects. Winds will pick up, moist places will dry and flames will roar.	Sci/Tech	Sports
Teenage T. rex's monster growth	Tyrannosaurus rex achieved its massive size due to an enormous growth spurt during its adolescent years.	Sci/Tech	Business
IBM to hire even more new workers	By the end of the year, the computing giant plans to have its biggest headcount since 1991.	Sci/Tech	Business
Some People Not Eligible to Get in on Google IPO	Google has billed its IPO as a way for everyday people to get in on the process, denying Wall Street the usual stranglehold it's had on IPOs. Public bidding, a minimum of just five shares, an open process with 28 underwriters - all this pointed to a new level of public participation. But this isn't the case.	Sci/Tech	Business
Rivals Try to Turn Tables on Charles Schwab	By MICHAEL LIEDTKE SAN FRANCISCO (AP) - With its low prices and iconoclastic attitude, discount stock broker Charles Schwab Corp. (SCH) represented an annoying stone in Wall Street's wing-tipped shoes for decades...	Sci/Tech	Business
Venezuela Prepares for Chavez Recall Vote	Supporters and rivals warn of possible fraud; government says Chavez's defeat could produce turmoil in world oil market.	World	Business
Live: Olympics day four	Richard Faulds and Stephen Parry are going for gold for Great Britain on day four in Athens.	World	Sports
Intel to delay product aimed for high-definition TVs	SAN FRANCISCO - In the latest of a series of product delays, Intel Corp. has postponed the launch of a video display chip it had previously planned to introduce by year end, putting off a showdown with Texas Instruments Inc. in the fast-growing market for high-definition television displays.	Business	Sci/Tech
U.S. Misses Cut in Olympic 100 Free	ATHENS, Greece - Top American sprinters Jason Lezak and Ian Crocker missed the cut in the Olympic 100-meter freestyle preliminaries Tuesday, a stunning blow for a country that had always done well in the event. Pieter van den Hoogenband of the Netherlands and Australian Ian Thorpe advanced to the evening semifinal a day after dueling teenager Michael Phelps in the 200 freestyle, won by Thorpe...	World	Sports
Stocks Climb on Drop in Consumer Prices	NEW YORK - Stocks rose for a second straight session Tuesday as a drop in consumer prices allowed investors to put aside worries about inflation, at least for the short term. With gasoline prices falling to eight-month lows, the Consumer Price Index registered a small drop in July, giving consumers a respite from soaring energy prices...	World	Business
Yahoo! Ups Ante for Small Businesses	Web giant Yahoo! is gambling that price cuts on its domain name registration and Web hosting products will make it more competitive with discounters in the space - which means that small businesses looking to move online get a sweeter deal through ...	Business	Sci/Tech
IBM Buys Two Danish Services Firms	IBM said Tuesday it has acquired a pair of Danish IT services firms as part of its effort to broaden its presence in Scandinavia. As a result of the moves, IBM will add about 3,700 IT staffers to its global head count. Financial terms of ...	Sci/Tech	Business
Oil prices bubble to record high	The price of oil has continued its sharp rise overnight, closing at a record high. The main contract in New York, light sweet crude for delivery next month, has closed at a record \$US46.75 a barrel - up 70 cents on yesterday's close.	World	Business
Security scare as intruder dives in	A CANADIAN husband's love for his wife has led to a tightening of security at all Olympic venues in Athens.	Sports	World
Google Lowers Its IPO Price Range	SAN JOSE, Calif. - In a sign that Google Inc.'s initial public offering isn't as popular as expected, the company lowered its estimated price range to between \$85 and \$95 per share, down from the earlier prediction of \$108 and \$135 per share...	World	Business
Future Doctors, Crossing Borders	Students at the Mount Sinai School of Medicine learn that diet and culture shape health in East Harlem.	World	Sci/Tech
Microsoft finalises three-year government deal	Hot on the heels of its 10-year strategic partnership with the London Borough of Newham, Microsoft is close to signing a new broad three-year public sector agreement with the government.	Sci/Tech	Business
Stock Prices Climb Ahead of Google IPO	NEW YORK - Investors shrugged off rising crude futures Wednesday to capture well-priced shares, sending the Nasdaq composite index up 1.6 percent ahead of Google Inc.'s much-anticipated initial public offering of stock. In afternoon trading, the Dow Jones industrial average gained 67.10, or 0.7 percent, to 10,039.93...	World	Business
Intuit Posts Wider Loss After Charge (Reuters)	Reuters - Intuit Inc. (INTU.O), maker of the No. 1 U.S. tax presentation software TurboTax, on Wednesday posted a wider quarterly loss after taking a goodwill impairment charge during its seasonally weaker fourth quarter.	Sci/Tech	Business
Hurricane center's projection on Charley not far off, data show	FORT LAUDERDALE, Fla. - (KRT) - Despite criticism that it should have better anticipated Hurricane Charley's rapid intensification and quick turn, the National Hurricane Center's forecast wasn't that far off, a preliminary post-mortem shows.	Sci/Tech	Business